

An Efficient Primal Dual Prox Method for Non-Smooth Optimization

Tianbao Yang

Mehrdad Mahdavi

Rong Jin

Department of Computer Science and Engineering

Michigan State University, East Lansing, MI, 48824, USA

Shenghuo Zhu

NEC Labs America, Cupertino, CA, 95014, USA

YANGTIA1@MSU.EDU

MAHDAVIM@MSU.EDU

RONGJIN@CSE.MSU.EDU

ZSH@SV.NEC-LABS.COM

Editor: ?

Abstract

We study the non-smooth optimization problems in machine learning, where both the loss function and the regularizer are non-smooth functions. Previous studies on efficient empirical loss minimization assume either a smooth loss function or a strongly convex regularizer, making them unsuitable for non-smooth optimization. We develop an efficient method for a family of non-smooth optimization where the dual form of the loss function is bilinear in primal and dual variables. We cast a non-smooth optimization problem into a minimax optimization problem, and develop a primal dual prox method that solves the minimax optimization problem at a rate of $O(1/T)$, significantly faster than a standard gradient descent method that has $O(1/\sqrt{T})$ convergence rate. Our empirical study verifies the efficiency of the proposed method for non-smooth optimization by comparing it to the state-of-the-art first order methods.

Keywords: non-smooth optimization, primal dual method, convergence rate, sparsity, efficiency

1. Introduction

Formulating machine learning tasks as a regularized empirical loss minimization problem makes an intimate connection between machine learning and mathematical optimization. In regularized empirical loss minimization, one tries to jointly minimize an empirical loss over training samples plus a regularization term of the model. This formulation includes support vector machine (SVM) (Hastie et al., 2008), support vector regression (Smola and Schölkopf, 2004), Lasso (Zhu et al., 2003), logistic regression, and ridge regression (Hastie et al., 2008) among many others. Therefore, optimization methods play a central role in solving machine learning problems and challenges exist in machine learning applications demand the development of new optimization algorithms.

Depending on the application at hand, various types of loss and regularization functions have been introduced in the literature. The efficiency of different optimization algorithms crucially depends on the specific structures of the loss and the regularization functions. Recently, there have been significant interests on gradient descent based methods due to

their simplicity and scalability to large datasets. A well-known example is the Pegasos algorithm (Shalev-Shwartz et al., 2007) which minimizes the ℓ_2^2 regularized hinge loss (i.e. SVM) and achieves a convergence rate of $O(1/T)$, where T is the number of iterations, by exploiting the strong convexity of the regularizer. Several other first order algorithms (Ji and Ye, 2009; Chen et al., 2009) are also proposed for smooth loss functions (e.g., squared loss and logistic loss) and non-smooth regularizers (i.e., $\ell_{1,\infty}$ and group lasso). They achieve a convergence rate of $O(1/T^2)$ by exploiting the smoothness of the loss functions.

In this paper, we focus on a more challenging case where both the loss function and the regularizer are non-smooth, to which we refer to as non-smooth optimization. Non-smooth optimization of regularized empirical loss has found applications in many machine learning problems. Examples of non-smooth loss functions include hinge loss (Vapnik, 1998), generalized hinge loss (Bartlett and Wegkamp, 2008), absolute loss (Hastie et al., 2008), and ϵ -insensitive loss (Rosasco et al., 2004); examples of non-smooth regularizers include lasso (Zhu et al., 2003), group lasso (Yuan and Lin, 2006), sparse group lasso (Yang et al., 2010), exclusive lasso (Zhou et al., 2010b), $\ell_{1,\infty}$ regularizer (Quattoni et al., 2009), and trace norm regularizer (Rennie and Srebro, 2005).

Although there are already many existing studies on tackling smooth loss functions (e.g. square loss for regression, logistic loss for classification), or smooth regularizers (e.g. ℓ_2^2 norm), there are serious challenges in developing efficient algorithms for non-smooth optimization. In particular, common tricks, such as smoothing non-smooth objective functions (Nesterov, 2005a,b), can not be applied to non-smooth optimization to improve convergence rate. This is because they require both the loss functions and regularizers be written in the maximization form of bilinear functions, which unfortunately are often violated, as we will discuss later. In this work, we focus on optimization problems in machine learning where both the loss function and the regularizer are non-smooth. Our goal is to develop an efficient gradient based algorithm that has a convergence rate of $O(1/T)$ for a wide family of non-smooth loss functions and general non-smooth regularizers.

It is noticeable that according to the information based complexity theory (Traub et al., 1988), it is impossible to derive an efficient first order algorithm that generally works for all non-smooth objective functions. As a result, we focus on a family of non-smooth optimization problems, where the dual form of the non-smooth loss function is bilinear in both primal and dual variables. Additionally, we show that many non-smooth loss functions have this bilinear dual form. We derive an efficient gradient based method, with a convergence rate of $O(1/T)$, that explicitly updates both the primal and dual variables. The proposed method is referred to as **Primal Dual Prox (Pdprox)** method. Besides its capability of dealing with non-smooth optimization, the proposed method is effective in handling the learning problems where additional constraints are introduced for dual variables.

The rest of this paper is organized as follows. Section 2 reviews the related work on minimizing regularized empirical loss especially the first order methods for large-scale optimization. Section 4 presents the proposed primal dual prox method, its convergence analysis, and several extensions of the proposed method. Section 5 presents the empirical studies, and Section 6 concludes this work.

2. Related Work

Our work is closely related to the previous studies on regularized empirical loss minimization. In the following discussion, we mostly focus on non-smooth loss functions and non-smooth regularizers.

Non-smooth loss functions Hinge loss is probably the most commonly used non-smooth loss function for classification. It is closely related to the max-margin criterion. A number of algorithms have been proposed to minimize the ℓ_2^2 regularized hinge loss (Platt, 1998; Joachims, 1999, 2006; Hsieh et al., 2008; Shalev-Shwartz et al., 2007), and the ℓ_1 regularized hinge loss (Cai et al., 2010; Zhu et al., 2003; Fung and Mangasarian, 2002). Besides the hinge loss, recently a generalized hinge loss function (Bartlett and Wegkamp, 2008) has been proposed for cost sensitive learning. For regression, square loss is commonly used due to its smoothness. However, non-smooth loss functions such as absolute loss (Hastie et al., 2008) and ϵ -insensitive loss (Rosasco et al., 2004) are useful for robust regression. The Bayes optimal predictor of square loss is the mean of the predictive distribution, while the Bayes optimal predictor of absolute loss is the median of the predictive distribution. Therefore absolute loss is more robust for long-tailed error distributions and outliers (Hastie et al., 2008). Rosasco et al. (2004) also proved that the estimation error bound for absolute loss and ϵ -insensitive loss converges faster than that of square loss. Non-smooth piecewise linear loss function has been used in quantile regression (Koenker, 2005; Gneiting, 2008). Unlike the absolute loss, the piecewise linear loss function can model non-symmetric error in reality.

Non-smooth regularizers Besides the simple non-smooth regularizers such as ℓ_1 , ℓ_2 , and ℓ_∞ norms (Duchi and Singer, 2009), many other non-smooth regularizers have been employed in machine learning tasks. Yuan and Lin (2006) introduced group lasso for selecting important explanatory factors in group manner. The $\ell_{1,\infty}$ norm regularizer has been used for multi-task learning (Argyriou et al., 2008). In addition, several recent works (Hou et al., 2011; Nie et al., 2010; Liu et al., 2009) considered mixed $\ell_{2,1}$ regularizer for feature selection. Zhou et al. (2010b) introduced exclusive lasso for multi-task feature selection to model the scenario where variables within a single group compete with each other. Trace norm regularizer is another non-smooth regularizer, which has found applications in matrix completion (Recht et al., 2010; Candès and Recht, 2008), matrix factorization (Rennie and Srebro, 2005; Srebro et al., 2005), and multi-task learning (Argyriou et al., 2008; Ji and Ye, 2009). The optimization algorithms presented in these works are usually limited: either the convergence rate is not guaranteed (Argyriou et al., 2008; Recht et al., 2010; Hou et al., 2011; Nie et al., 2010; Rennie and Srebro, 2005; Srebro et al., 2005) or the loss functions are assumed to be smooth (e.g., the square loss or the logistic loss) (Liu et al., 2009; Ji and Ye, 2009). Despite the significant efforts in developing algorithms for minimizing regularized empirical losses, it remains a challenge to design a first order algorithm that is able to efficiently solve non-smooth optimization problems at a rate of $O(1/T)$ when both the loss function and the regularizer are non-smooth.

Gradient based optimization Our work is closely related to (sub)gradient based optimization methods. The convergence rate of gradient based methods usually depends on the properties of the objective function to be optimized. When the objective function is strongly convex and smooth, it is well known that gradient descent methods can achieve

a geometric convergence rate (Boyd and Vandenberghe, 2004). When the objective function is smooth but not strongly convex, the optimal convergence rate of a gradient descent method is $O(1/T^2)$, and is achieved by the Nesterov’s methods (Nesterov, 2007). For the objective function which is strongly convex but not smooth, the convergence rate becomes $O(1/T)$ (Shalev-Shwartz et al., 2007). For general non-smooth objective functions, the optimal rate of any first order method is $O(1/\sqrt{T})$. Although it is not improvable in general, recent studies are able to improve this rate to $O(1/T)$ by exploring the special structure of the objective function (Nesterov, 2005a,b). In addition, several methods are developed for composite optimization, where the objective function is written as a sum of a smooth and a non-smooth function (Lan, 2010; Nesterov, 2007; Lin, 2010). Recently, these optimization techniques have been successfully applied to various machine learning problems, such as SVM (Zhou et al., 2010a), general regularized empirical loss minimization (Duchi and Singer, 2009; Hu et al., 2009), trace norm minimization (Ji and Ye, 2009), and multi-task sparse learning (Chen et al., 2009). Despite these efforts, one major limitation of the existing (sub)gradient based algorithms is that in order to achieve a convergence rate better than $O(1/\sqrt{T})$, they have to assume that the loss function is smooth or the regularizer is strongly convex, making them unsuitable for non-smooth optimization.

Convex-concave optimization The present work is also related to convex-concave minimization. Tseng (2008) and Nemirovski (2005) developed prox methods that have a convergence rate of $O(1/T)$, provided the gradients are Lipschitz continuous and have been applied to machine learning problems (Sun et al., 2009). In contrast, our method achieves a rate of $O(1/T)$ without requiring gradients to be Lipschitz continuous. Several other primal-dual algorithms have been developed for regularized empirical loss minimization that update both primal and dual variables. Zhu and Chan (2008) propose a primal-dual method based on gradient descent, which only achieves a rate of $O(1/\sqrt{T})$. Mosci et al. (2010) present a primal-dual algorithm for group sparse regularization, which updates the primal variable by a prox method and the dual variable by a Newton’s method. In contrast, the proposed algorithm is a first order method that does not require computing the Hessian matrix as the Newton’s method does, and is therefore more scalable to large datasets. Lan et al. (2011) consider the primal-dual convex formulations for general cone programming and apply Nesterov’s optimal first order method (Nesterov, 2007), Nesterov’s smoothing technique (Nesterov, 2005a), and Nemirovski’s prox method (Nemirovski, 2005). Nesterov (2005b) proposed a primal dual gradient method for a special class of structured non-smooth optimization problems by exploring an excessive gap technique.

Optimizing non-smooth functions We note that Nesterov’s smoothing technique (Nesterov, 2005a) and excessive gap technique (Nesterov, 2005b) can be applied to non-smooth optimization and both achieve $O(1/T)$ convergence rate for a special class of non-smooth optimization problems. However, the limitation of these approaches is that they require all the non-smooth terms (i.e., the loss and the regularizer) to be written as an explicit max structure that consists of a bilinear function in primal and dual variables, thus limits their applications to many machine learning problems. In addition, Nesterov’s algorithms need to solve additional maximizations problem at each iteration. In contrast, the proposed algorithm only requires mild condition on the non-smooth loss functions (section 4), and allows for any commonly used non-smooth regularizers, without having to solve an

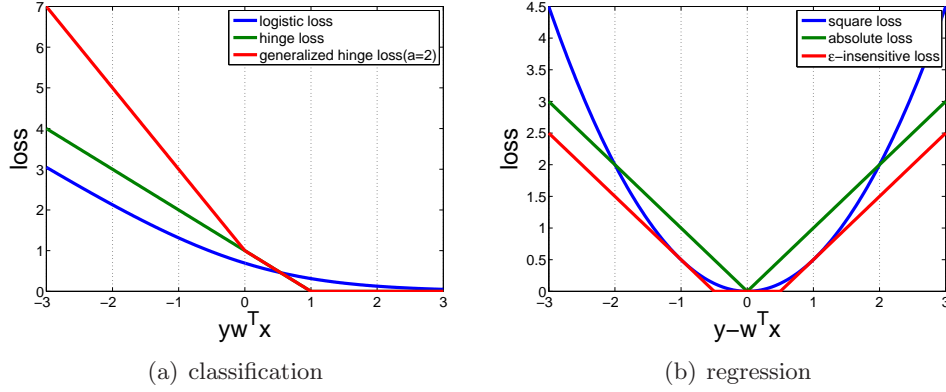


Figure 1: Loss functions

additional optimization problem at each iteration. Compared to Nesterov's algorithms, the proposed algorithm is applicable to a large class of non-smooth optimization problems, is easier to implement, its convergence analysis is much simpler, and its empirical performance is usually better.

3. Notations and Definitions

In this section we provide the basic setup, some preliminary definitions and notations used throughout this paper.

We denote by $[n]$ the set of integers $\{1, \dots, n\}$. We denote by $(\mathbf{x}_i, y_i), i \in [n]$ the training examples, where $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^d$ and y_i is the assigned class label, which is discrete for classification and continuous for regression. We assume $\|\mathbf{x}_i\|_2 \leq 1, \forall i \in [n]$. We denote by $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$ and $\mathbf{y} = (y_1, \dots, y_n)^\top$. Let $\mathbf{w} \in \mathbb{R}^d$ denote the linear hypothesis, $\ell(\mathbf{w}; \mathbf{x}, y)$ denote a loss of prediction made by the hypothesis \mathbf{w} on example (\mathbf{x}, y) , which is a convex function in terms of \mathbf{w} . Examples of convex loss function are hinge loss $\ell(\mathbf{w}; \mathbf{x}, y) = \max(1 - y\mathbf{w}^\top \mathbf{x}, 0)$, and absolute loss $\ell(\mathbf{w}; \mathbf{x}, y) = |\mathbf{w}^\top \mathbf{x} - y|$. To characterize a function, we introduce the following definitions

Definition 1 A function $\ell(\mathbf{z}) : \mathcal{Z} \rightarrow \mathbb{R}$ is a G -Lipschitz continuous if

$$|\ell(\mathbf{z}_1) - \ell(\mathbf{z}_2)| \leq G\|\mathbf{z}_1 - \mathbf{z}_2\|_2, \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}.$$

Definition 2 A function $\ell(\mathbf{z}) : \mathcal{Z} \rightarrow \mathbb{R}$ is a ρ -smooth function if its gradient is ρ -Lipschitz continuous

$$\|\nabla \ell(\mathbf{z}_1) - \nabla \ell(\mathbf{z}_2)\|_2 \leq \rho\|\mathbf{z}_1 - \mathbf{z}_2\|_2, \forall \mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}.$$

A function is non-smooth if either its gradient is not well defined or its gradient is not Lipschitz continuous. Examples of smooth loss functions are logistic loss $\ell(\mathbf{w}; \mathbf{x}, y) = \log(1 + \exp(-y\mathbf{w}^\top \mathbf{x}))$, square loss $\ell(\mathbf{w}; \mathbf{x}, y) = \frac{1}{2}(\mathbf{w}^\top \mathbf{x} - y)^2$, and examples of non-smooth loss functions are hinge loss, and absolute loss. The difference between logistic loss and hinge loss, square loss and absolute loss can be seen in Figure 1. Examples of non-smooth

regularizer include $R(\mathbf{w}) = \|\mathbf{w}\|_1$, i.e. ℓ_1 norm, $R(\mathbf{w}) = \|\mathbf{w}\|_\infty$, i.e. ℓ_∞ norm. More examples can be found in section 4.1.

In this paper, we aim to solve the following optimization problem, which occurs in many machine learning problems,

$$\min_{\mathbf{w} \in \mathbb{R}^d} \mathcal{L}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{w}; \mathbf{x}_i, y_i) + \lambda R(\mathbf{w}), \quad (1)$$

where $\ell(\mathbf{w}; \mathbf{x}, y)$ is a non-smooth loss function, $R(\mathbf{w})$ is a non-smooth regularizer on \mathbf{w} , and λ is a regularization parameter.

We denote by $\Pi_{\mathcal{Q}}[\hat{\mathbf{z}}] = \arg \min_{\mathbf{z} \in \mathcal{Q}} \frac{1}{2} \|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$ the projection of $\hat{\mathbf{z}}$ into domain \mathcal{Q} , and by $\Pi_{\mathcal{Q}_1, \mathcal{Q}_2} \begin{pmatrix} \hat{\mathbf{z}}_1 \\ \hat{\mathbf{z}}_2 \end{pmatrix}$ the joint projection of $\hat{\mathbf{z}}_1$ and $\hat{\mathbf{z}}_2$ into domains \mathcal{Q}_1 and \mathcal{Q}_2 , respectively. Finally, we use $[s]_{[0, s_+]}$ to denote the projection of s into $[0, s_+]$.

4. Pdprox: A Primal Dual Prox Method for Non-Smooth Optimization

We first describe the non-smooth optimization problems that the proposed algorithm can be applied to, and then present the primal dual prox method for non-smooth optimization. We prove the convergence bound of the proposed algorithms and discuss several extensions. Proofs for technical lemmas are deferred to the appendix.

4.1 Non-Smooth Optimization

We first focus our analysis on linear classifiers and denote by $\mathbf{w} \in \mathbb{R}^d$ a linear model. The extension to nonlinear models is discussed in section 4.5. Also, extension to a collection of linear models $\mathbf{W} \in \mathbb{R}^{d \times K}$ can be done in a straightforward way. We consider the following general *non-smooth* optimization problem:

$$\min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} \left[\mathcal{L}(\mathbf{w}) = \max_{\alpha \in \mathcal{Q}_{\alpha}} L(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y}) + \lambda R(\mathbf{w}) \right]. \quad (2)$$

The parameters \mathbf{w} in domain $\mathcal{Q}_{\mathbf{w}}$ and α in domain \mathcal{Q}_{α} are referred to as primal and dual variables, respectively. Since it is impossible to develop an efficient first order method for general non-smooth optimization, we focus on the family of non-smooth loss functions that can be characterized by bilinear function $L(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y})$, i.e.

$$L(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y}) = c_0(\mathbf{X}, \mathbf{y}) + \alpha^\top \mathbf{a}(\mathbf{X}, \mathbf{y}) + \mathbf{w}^\top \mathbf{b}(\mathbf{X}, \mathbf{y}) + \mathbf{w}^\top \mathbf{H}(\mathbf{X}, \mathbf{y}) \alpha, \quad (3)$$

where $c_0(\mathbf{X}, \mathbf{y})$, $\mathbf{a}(\mathbf{X}, \mathbf{y})$, $\mathbf{b}(\mathbf{X}, \mathbf{y})$, and $\mathbf{H}(\mathbf{X}, \mathbf{y})$ are the parameters depending on the training examples (\mathbf{X}, \mathbf{y}) with consistent sizes. In the sequel, we denote by $L(\mathbf{w}, \alpha) = L(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y})$ for simplicity, and by $G_{\mathbf{w}}(\mathbf{w}, \alpha) = \nabla_{\mathbf{w}} L(\mathbf{w}, \alpha)$ and $G_{\alpha}(\mathbf{w}, \alpha) = \nabla_{\alpha} L(\mathbf{w}, \alpha)$ the partial gradients of $L(\mathbf{w}, \alpha)$ in terms of \mathbf{w} and α , respectively.

Remark 1 One direct consequence of assumption in (3) is that the partial gradient $G_{\mathbf{w}}(\mathbf{w}, \alpha)$ is independent of \mathbf{w} , and $G_{\alpha}(\mathbf{w}, \alpha)$ is independent of α , since $L(\mathbf{w}, \alpha)$ is bilinear in \mathbf{w} and α . We will explicitly exploit this property in developing the efficient optimization

algorithms. We also note that no explicit assumption is made for the regularizer $R(\mathbf{w})$. This is in contrast to the smoothing techniques used in (Nesterov, 2005a,b).

To efficiently solve the optimization problem in (1), we need first turn it into the form (2). To this end, we assume that the loss function can be written into a dual form, which is bilinear in the primal and the dual variables, i.e.

$$\ell(\mathbf{w}; \mathbf{x}_i, y_i) = \max_{\alpha_i \in \Delta_\alpha} f(\mathbf{w}, \alpha_i; \mathbf{x}_i, y_i), \quad (4)$$

where $f(\mathbf{w}, \alpha; \mathbf{x}, y)$ is a bilinear function in \mathbf{w} and α , and Δ_α is the domain of variable α . Using (4), we cast problem (1) into (2) with $L(\mathbf{w}, \boldsymbol{\alpha}; \mathbf{X}, \mathbf{y})$ given by

$$L(\mathbf{w}, \boldsymbol{\alpha}; \mathbf{X}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n f(\mathbf{w}, \alpha_i; \mathbf{x}_i, y_i), \quad (5)$$

with $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top$ defined in the domain $\mathcal{Q}_\alpha = \{\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top, \alpha_i \in \Delta_\alpha\}$.

Before delving into the description of the proposed algorithms and their analysis, we give a few examples that shows many non-smooth loss functions can be written in the form of (4):

- Hinge loss (Vapnik, 1998):

$$\ell(\mathbf{w}; \mathbf{x}, y) = \max(0, 1 - y\mathbf{w}^\top \mathbf{x}) = \max_{\alpha \in [0,1]} \alpha(1 - y\mathbf{w}^\top \mathbf{x}).$$

- Generalized hinge loss (Bartlett and Wegkamp, 2008):

$$\begin{aligned} \ell(\mathbf{w}; \mathbf{x}, y) &= \begin{cases} 1 - ay\mathbf{w}^\top \mathbf{x} & \text{if } y\mathbf{w}^\top \mathbf{x} \leq 0 \\ 1 - y\mathbf{w}^\top \mathbf{x} & \text{if } 0 < y\mathbf{w}^\top \mathbf{x} < 1 \\ 0 & \text{if } y\mathbf{w}^\top \mathbf{x} \geq 1 \end{cases} \\ &= \max_{\substack{\alpha_1 \geq 0, \alpha_2 \geq 0 \\ \alpha_1 + \alpha_2 \leq 1}} \alpha_1(1 - ay\mathbf{w}^\top \mathbf{x}) + \alpha_2(1 - y\mathbf{w}^\top \mathbf{x}), \end{aligned}$$

where $a > 1$.

- Absolute loss (Hastie et al., 2008):

$$\ell(\mathbf{w}; \mathbf{x}, y) = |\mathbf{w}^\top \mathbf{x} - y| = \max_{\alpha \in [-1,1]} \alpha(\mathbf{w}^\top \mathbf{x} - y).$$

- ϵ -insensitive loss (Rosasco et al., 2004) :

$$\ell(\mathbf{w}; \mathbf{x}, y) = \max(|\mathbf{w}^\top \mathbf{x} - y| - \epsilon, 0) = \max_{\substack{\alpha_1 \geq 0, \alpha_2 \geq 0 \\ \alpha_1 + \alpha_2 \leq 1}} \left[(\mathbf{w}^\top \mathbf{x} - y)(\alpha_1 - \alpha_2) - \epsilon(\alpha_1 + \alpha_2) \right].$$

- Piecewise linear loss (Koenker, 2005):

$$\begin{aligned} \ell(\mathbf{w}; \mathbf{x}, y) &= \begin{cases} a|\mathbf{w}^\top \mathbf{x} - y| & \text{if } \mathbf{w}^\top \mathbf{x} \leq y \\ (1-a)|\mathbf{w}^\top \mathbf{x} - y| & \text{if } \mathbf{w}^\top \mathbf{x} \geq y \end{cases} \\ &= \max_{\substack{\alpha_1 \geq 0, \alpha_2 \geq 0 \\ \alpha_1 + \alpha_2 \leq 1}} \alpha_1 a(y - \mathbf{w}^\top \mathbf{x}) + \alpha_2 (1-a)(\mathbf{w}^\top \mathbf{x} - y). \end{aligned}$$

- ℓ_2 loss (Nie et al., 2010):

$$\ell(\mathbf{W}; \mathbf{x}, \mathbf{y}) = \|\mathbf{W}^\top \mathbf{x} - \mathbf{y}\|_2 = \max_{\|\alpha\|_2 \leq 1} \alpha^\top (\mathbf{W}^\top \mathbf{x} - \mathbf{y}),$$

where $\mathbf{y} \in \mathbb{R}^K$ is multiple class label vector and $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$.

Besides the non-smooth loss function $\ell(\mathbf{w}; \mathbf{x}, y)$, we also assume that the regularizer $R(\mathbf{w})$ is a non-smooth function. Many non-smooth regularizers are used in machine learning problems. We list a few of them in the following, where $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_K)$, $\mathbf{w}_k \in \mathbb{R}^d$ and \mathbf{w}^j is the j th row of \mathbf{W} .

- lasso: $R(\mathbf{w}) = \|\mathbf{w}\|_1$, ℓ_2 norm: $R(\mathbf{w}) = \|\mathbf{w}\|_2$, and ℓ_∞ norm: $R(\mathbf{w}) = \|\mathbf{w}\|_\infty$.
- group lasso: $R(\mathbf{w}) = \sum_{g=1}^K \sqrt{d_g} \|\mathbf{w}_g\|_2$, where $\mathbf{w}_g \in \mathbb{R}^{d_g}$.
- exclusive lasso: $R(\mathbf{W}) = \sum_{j=1}^d \|\mathbf{w}^j\|_1^2$.
- $\ell_{2,1}$ norm: $R(\mathbf{W}) = \sum_{j=1}^d \|\mathbf{w}^j\|_2$.
- $\ell_{1,\infty}$ norm: $R(\mathbf{W}) = \sum_{j=1}^d \|\mathbf{w}^j\|_\infty$.
- trace norm: $R(\mathbf{W}) = \|\mathbf{W}\|_1$, the summation of singular values of \mathbf{W} .
- other regularizers: $R(\mathbf{W}) = \left(\sum_{k=1}^K \|\mathbf{w}_k\|_2 \right)^2$.

Note that unlike (Nesterov, 2005a,b), we do not further require the non-smooth regularizer to be written into a bilinear dual form, which could be violated by many non-smooth regularizers, e.g. $R(\mathbf{W}) = \left(\sum_{k=1}^K \|\mathbf{w}_k\|_2 \right)^2$ or more generally $R(\mathbf{w}) = V(\|\mathbf{w}\|)$, where $V(z)$ is a monotonically increasing function.

We end this section by presenting a lemma showing an important property of the bilinear function $L(\mathbf{w}, \alpha)$.

Lemma 3 *Let $L(\mathbf{w}, \alpha)$ be bilinear in \mathbf{w} and α as in (3). Assuming there exists $c > 0$ such that $\|H(\mathbf{X}, \mathbf{y})\|_2^2 \leq c$, then for any $\alpha_1, \alpha_2 \in \mathcal{Q}_\alpha$, and $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{Q}_\mathbf{w}$ we have*

$$\|G_\alpha(\mathbf{w}_1, \alpha_1) - G_\alpha(\mathbf{w}_2, \alpha_2)\|_2^2 \leq c \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \quad (6)$$

$$\|G_\mathbf{w}(\mathbf{w}_1, \alpha_1) - G_\mathbf{w}(\mathbf{w}_2, \alpha_2)\|_2^2 \leq c \|\alpha_1 - \alpha_2\|_2^2. \quad (7)$$

Remark 2 The value of constant c in Lemma 3 is an input to our algorithms used to set the step size. In the Appendix A, we show how to estimate constant c for certain loss functions. In addition the constant c in bounds (6) and (7) do not have to be the same as shown by the example of generalized hinge loss in Appendix A.

Algorithm 1 The Pduprox-dual Algorithm for Non-Smooth Optimization

- 1: **Input:** step size $\gamma = \sqrt{1/(2c)}$, where c is specified in (6).
 - 2: **Initialization:** $\mathbf{w}_0 = \mathbf{0}, \beta_0 = \mathbf{0}$
 - 3: **for** $t = 1, 2, \dots$ **do**
 - 4: $\alpha_t = \Pi_{Q_\alpha} [\beta_{t-1} + \gamma G_\alpha(\mathbf{w}_{t-1}, \beta_{t-1})]$
 - 5: $\mathbf{w}_t = \arg \min_{\mathbf{w} \in Q_w} \frac{1}{2} \|\mathbf{w} - (\mathbf{w}_{t-1} - \gamma G_w(\mathbf{w}_{t-1}, \alpha_t))\|_2^2 + \gamma \lambda R(\mathbf{w})$
 - 6: $\beta_t = \Pi_{Q_\alpha} [\beta_{t-1} + \gamma G_\alpha(\mathbf{w}_t, \alpha_t)]$
 - 7: **end for**
 - 8: **Output** $\hat{\mathbf{w}}_T = \sum_{t=1}^T \mathbf{w}_t / T$ and $\hat{\alpha}_T = \sum_{t=1}^T \alpha_t / T$.
-

4.2 The Proposed Primal-Dual Prox Methods

In this subsection, we present two variants of Primal Dual Prox (Pduprox) method for solving the non-smooth optimization problem in (2). The common feature shared by the two algorithms is that they update both the primal and the dual variables at each iteration. In contrast, most first order methods only update the primal variables. The key advantages of the proposed algorithms is that they are able to capture the **sparsity structures of both primal and dual variables**, which is usually the case when both the regularizer and the loss functions are both non-smooth. The two algorithms differ from each other in the number of copies for the dual or the primal variables, and the specific order for updating those. Although our analysis shows that the two algorithms share the same convergence rate; however, our empirical studies show that the second algorithm is more advantageous when the number of training examples is significantly larger than the number of features (i.e. $n \gg d$).

Pduprox-dual algorithm Algorithm 1 shows the first primal dual prox algorithm for optimizing the problem in (2). Compared to the other gradient based algorithms, Algorithm 1 has several interesting features:

- (i) it updates both the dual variable α and the primal variable \mathbf{w} . This is useful when additional constraints are introduced for the dual variables, as we will discuss later.
- (ii) it introduces an auxiliary dual variable β in addition to α , and updates both α and β at each iteration by a gradient mapping. The gradient mapping on the dual variables into a sparse domain allows the proposed algorithm to capture the sparsity of the dual variables (more discussion on how the sparse constraint on the dual variable affects the convergence is presented in Section 4.5). Compared to the second algorithm presented below, we refer to Algorithm 1 as **Pduprox-dual** algorithm since it introduces an auxiliary dual variable in updating.
- (iii) the primal variable \mathbf{w} is updated by a composite gradient mapping (Nesterov, 2007) in step 5. Solving a composite gradient mapping in this step allows the proposed algorithm to capture the sparsity of the primal variable. Similar to many other approaches for composite optimization (Duchi and Singer, 2009; Hu et al., 2009), we assume that the mapping in step 5 can be solved efficiently. (This is the only assumption we made on the non-smooth regularizer. The discussion in Section 4.4 shows that the proposed algorithm can be applied to a large family of non-smooth regularizers).

Algorithm 2 The Pdprox-primal Algorithm for Non-Smooth Optimization

- 1: **Input:** step size $\gamma = \sqrt{1/(2c)}$, where c is specified in (7).
 - 2: **Initialization:** $\mathbf{u}_0 = \mathbf{0}, \boldsymbol{\alpha}_0 = \mathbf{0}$
 - 3: **for** $t = 1, 2, \dots$ **do**
 - 4: $\mathbf{w}_t = \arg \min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} \frac{1}{2} \|\mathbf{w} - (\mathbf{u}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1}))\|_2^2 + \gamma \lambda R(\mathbf{w})$
 - 5: $\boldsymbol{\alpha}_t = \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}} [\boldsymbol{\alpha}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_{t-1})]$
 - 6: $\mathbf{u}_t = \mathbf{w}_t + \gamma (G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1}) - G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t))$
 - 7: **end for**
 - 8: **Output** $\hat{\mathbf{w}}_T = \sum_{t=1}^T \mathbf{w}_t / T$ and $\hat{\boldsymbol{\alpha}}_T = \sum_{t=1}^T \boldsymbol{\alpha}_t / T$.
-

- (iv) the step size γ is fixed to $\sqrt{1/(2c)}$, where c is the constant specified in Lemma 3. This is in contrast to most gradient based methods where the step size depends on T and/or λ . This feature is particularly useful in implementation as we often observe that the performance of a gradient method is sensitive to the choice of the step size.

Pdprox-primal algorithm In Algorithm 1, we maintain two copies of the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and update them by two gradient mappings. We can actually save one gradient mapping on the dual variable by first updating the primal variable \mathbf{w}_t , and then updating $\boldsymbol{\alpha}_t$ using partial gradient computed with \mathbf{w}_t . As a tradeoff, we add an auxiliary primal variable \mathbf{u} , and update it by a simple calculation. The detailed steps are shown in Algorithm 2. Similar to Algorithm 1, Algorithm 2 also needs to compute two partial gradients (except the initial partial gradient on the primal variable), i.e. $G_{\mathbf{w}}(\cdot, \boldsymbol{\alpha}_t)$ and $G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \cdot)$. Different from Algorithm 1, Algorithm 2 (i) maintains $(\mathbf{w}_t, \boldsymbol{\alpha}_t, \mathbf{u}_t)$ at each iteration with $O(2d + n)$ memory, while Algorithm 1 maintains $(\boldsymbol{\alpha}_t, \mathbf{w}_t, \boldsymbol{\beta}_t)$ at each iteration with $O(2n + d)$ memory; (ii) and replaces one gradient mapping on an auxiliary dual variable $\boldsymbol{\beta}_t$ with a simple update on an auxiliary primal variable \mathbf{u}_t . When the number of examples n is much larger than the dimension d , i.e. $n \gg d$, Algorithm 2 could save the memory and the computational cost. However, the convergence rate of Algorithm 2 is the same to Algorithm 1. Because it introduces an auxiliary primal variable, we refer to Algorithm 2 as the **Pdprox-primal** algorithm.

It should be noted that although Algorithm 1 uses a similar strategy for updating the dual variables $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, but it is significantly different from the mirror prox method (Nemirovski, 2005). First, unlike the mirror prox method that introduces an auxiliary variable for \mathbf{w} , Algorithm 1 introduces a composite gradient mapping for updating \mathbf{w} . Second, Algorithm 1 updates \mathbf{w}_t using the partial gradient computed from the updated dual variable $\boldsymbol{\alpha}_t$ rather than $\boldsymbol{\beta}_{t-1}$. Third, Algorithm 1 does not assume that the overall objective function has Lipschitz continuous gradients, a key assumption that limits the application of the mirror prox method.

4.3 Convergence Analysis

This section establishes bounds on the convergence rate of the proposed algorithms. We begin by presenting a theorem about the convergence rate of Algorithms 1 and 2. For ease

of analysis, we first write (2) into the following equivalent minimax formulation

$$\min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} \max_{\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}} F(\mathbf{w}, \boldsymbol{\alpha}) = L(\mathbf{w}, \boldsymbol{\alpha}) + \lambda R(\mathbf{w}). \quad (8)$$

Our main result is stated in the following theorem:

Theorem 4 *By running Algorithm 1 or Algorithm 2 with T steps, we have*

$$F(\hat{\mathbf{w}}_T, \boldsymbol{\alpha}) - F(\mathbf{w}, \hat{\boldsymbol{\alpha}}_T) \leq \frac{\|\mathbf{w}\|_2^2 + \|\boldsymbol{\alpha}\|_2^2}{\sqrt{(2/c)T}},$$

for any $\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}$ and $\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}$. In particular,

$$\max_{\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}} F(\hat{\mathbf{w}}_T, \boldsymbol{\alpha}) - \min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} F(\mathbf{w}, \hat{\boldsymbol{\alpha}}_T) \leq \frac{\|\mathbf{w}^*\|_2^2 + \|\boldsymbol{\alpha}^*\|_2^2}{\sqrt{(2/c)T}},$$

where $(\mathbf{w}^*, \boldsymbol{\alpha}^*)$ is the optimal solution to $\max_{\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}} F(\hat{\mathbf{w}}_T, \boldsymbol{\alpha}) - \min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} F(\mathbf{w}, \hat{\boldsymbol{\alpha}}_T)$.

In order to aid understanding, we present the proof of Theorem 4 for each algorithm separately in the following subsections.

4.3.1 CONVERGENCE ANALYSIS OF ALGORITHM 1

For the simplicity of analysis, we assume $\mathcal{Q}_{\mathbf{w}} = \mathbb{R}^d$ is the whole Euclidean space. We discuss how to generalize the analysis to a convex domain $\mathcal{Q}_{\mathbf{w}}$ in Section 4.5. In order to prove Theorem 4 for Algorithm 1, we present a series of lemmas to pave the path for the proof. We first restate the key updates in Algorithm 1 as follows:

$$\boldsymbol{\alpha}_t = \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}} [\boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_{t-1}, \boldsymbol{\beta}_{t-1})], \quad (9)$$

$$\mathbf{w}_t = \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w} - (\mathbf{w}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{w}_{t-1}, \boldsymbol{\alpha}_t))\|_2^2 + \gamma \lambda R(\mathbf{w}), \quad (10)$$

$$\boldsymbol{\beta}_t = \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}} [\boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t)]. \quad (11)$$

Lemma 5 *The updates in Algorithm 1 are equivalent to the following gradient mappings,*

$$\begin{pmatrix} \boldsymbol{\alpha}_t \\ \mathbf{w}_t \end{pmatrix} = \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}, \mathbb{R}^d} \begin{pmatrix} \boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1}) \\ \mathbf{u}_{t-1} - \gamma(G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t) \end{pmatrix},$$

and

$$\begin{pmatrix} \boldsymbol{\beta}_t \\ \mathbf{u}_t \end{pmatrix} = \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}, \mathbb{R}^d} \begin{pmatrix} \boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) \\ \mathbf{u}_{t-1} - \gamma(G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t) \end{pmatrix},$$

with initialization $\mathbf{u}_0 = \mathbf{w}_0$, where $\mathbf{v}_t \in \partial R(\mathbf{w}_t)$ is a partial gradient of the regularizer on \mathbf{w}_t .

Proof First, we argue that there exists a fixed (sub)gradient $\mathbf{v}_t \in \partial R(\mathbf{w}_t)$ such that the composite gradient mapping (10) is equivalent to the following gradient mapping,

$$\mathbf{w}_t = \Pi_{\mathbb{R}^d} [\mathbf{w}_{t-1} - \gamma (G_{\mathbf{w}}(\mathbf{w}_{t-1}, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t)]. \quad (12)$$

To see this, since \mathbf{w}_t is the optimal solution to (10), by first order optimality condition, there exists a subgradient $\mathbf{v}_t = \partial R(\mathbf{w}_t)$ such that $\mathbf{w}_t - \mathbf{w}_{t-1} + \gamma G_{\mathbf{w}}(\mathbf{w}_{t-1}, \boldsymbol{\alpha}_t) + \gamma \lambda \mathbf{v}_t = \mathbf{0}$, i.e.

$$\mathbf{w}_t = \mathbf{w}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{w}_{t-1}, \boldsymbol{\alpha}_t) - \gamma \lambda \mathbf{v}_t,$$

which is equivalent to (12) since the projection $\Pi_{\mathbb{R}^d}$ is an identical mapping.

Second, the updates in Algorithm 1 for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{u})$ are equivalent to the following updates for $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{w}, \mathbf{u})$

$$\begin{aligned} \boldsymbol{\alpha}_t &= \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}} [\boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1})], \\ \mathbf{w}_t &= \Pi_{\mathbb{R}^d} [\mathbf{u}_{t-1} - \gamma (G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t)], \end{aligned} \quad (13)$$

$$\begin{aligned} \boldsymbol{\beta}_t &= \Pi_{\mathcal{Q}_{\boldsymbol{\alpha}}} [\boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t)], \\ \mathbf{u}_t &= \mathbf{w}_t - \gamma (G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) - G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_t)), \end{aligned} \quad (14)$$

with initialization $\mathbf{u}_0 = \mathbf{w}_0$. The reason is because $\mathbf{u}_t = \mathbf{w}_t, t = 1, \dots$ due to $G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) = G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_t)$, where we use the fact that $L(\mathbf{w}, \boldsymbol{\alpha})$ is linear in \mathbf{w} .

Finally, by plugging (13) for \mathbf{w}_t into the update for \mathbf{u}_t in (14), we complete the proof of Lemma 5. \blacksquare

The reason that we translate the updates for $(\boldsymbol{\alpha}_t, \mathbf{w}_t, \boldsymbol{\beta}_t)$ in Algorithm 1 into the updates for $(\boldsymbol{\alpha}_t, \mathbf{w}_t, \boldsymbol{\beta}_t, \mathbf{u}_t)$ in Lemma 5 is because it allows us to fit the updates for $(\boldsymbol{\alpha}_t, \mathbf{w}_t, \boldsymbol{\beta}_t, \mathbf{u}_t)$ into Lemma 12 as presented in Appendix C, which leads us to a key inequality as stated in Lemma 6 to prove Theorem 4.

Lemma 6 For all $t = 1, 2, \dots$, and any $\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}$, we have

$$\begin{aligned} \gamma \begin{pmatrix} G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t \\ -G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) \end{pmatrix}^\top \begin{pmatrix} \mathbf{w}_t - \mathbf{w} \\ \boldsymbol{\alpha}_t - \boldsymbol{\alpha} \end{pmatrix} &\leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2 \\ &\quad + \gamma^2 \left\| G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) - G_{\boldsymbol{\alpha}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1}) \right\|_2^2 - \frac{1}{2} \|\mathbf{w}_t - \mathbf{u}_{t-1}\|_2^2. \end{aligned}$$

The proof of Lemma 6 is deferred to Appendix C. We are now ready to prove the main theorem for Algorithm 1.

Proof [of Theorem 4 for Algorithm 1] Since $F(\mathbf{w}, \boldsymbol{\alpha})$ is convex in \mathbf{w} and concave in $\boldsymbol{\alpha}$, we have

$$\begin{aligned} F(\mathbf{w}_t, \boldsymbol{\alpha}_t) - F(\mathbf{w}, \boldsymbol{\alpha}_t) &\leq (G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t)^\top (\mathbf{w}_t - \mathbf{w}), \\ F(\mathbf{w}_t, \boldsymbol{\alpha}) - F(\mathbf{w}_t, \boldsymbol{\alpha}_t) &\leq -G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t)^\top (\boldsymbol{\alpha}_t - \boldsymbol{\alpha}), \end{aligned}$$

where $\mathbf{v}_t \in \partial R(\mathbf{w}_t)$ is the partial gradient of $R(\mathbf{w})$ on \mathbf{w}_t stated in Lemma 5. Combining the above inequalities with Lemma 6, we have

$$\begin{aligned}
 & \gamma (F(\mathbf{w}_t, \boldsymbol{\alpha}_t) - F(\mathbf{w}, \boldsymbol{\alpha}_t) + F(\mathbf{w}_t, \boldsymbol{\alpha}) - F(\mathbf{w}_t, \boldsymbol{\alpha}_t)) \\
 & \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2 + \gamma^2 \|G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) - G_{\boldsymbol{\alpha}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1})\|_2^2 - \frac{1}{2} \|\mathbf{w}_t - \mathbf{u}_{t-1}\|_2^2 \\
 & \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2 + \gamma^2 c \|\mathbf{w}_t - \mathbf{u}_{t-1}\|_2^2 - \frac{1}{2} \|\mathbf{w}_t - \mathbf{u}_{t-1}\|_2^2 \\
 & \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2,
 \end{aligned}$$

where the second inequality follows the inequality (6) in Lemma 3 and the fact $\gamma = \sqrt{1/(2c)}$. By adding the inequalities of all iterations, we have

$$\sum_{t=1}^T (F(\mathbf{w}_t, \boldsymbol{\alpha}) - F(\mathbf{w}, \boldsymbol{\alpha}_t)) \leq \frac{\|\mathbf{w}\|_2^2 + \|\boldsymbol{\alpha}\|_2^2}{\sqrt{(2/c)} T}. \quad (15)$$

We complete the proof by using the definitions of $\widehat{\mathbf{w}}_T, \widehat{\boldsymbol{\alpha}}_T$, and the convexity-concavity of $F(\mathbf{w}, \boldsymbol{\alpha})$ with respect to \mathbf{w} and $\boldsymbol{\alpha}$, respectively. \blacksquare

4.3.2 CONVERGENCE ANALYSIS OF ALGORITHM 2

We can prove the convergence bound for Algorithm 2 by following the same path. In the following we present the key lemmas similar to Lemmas 5 and 6, with proofs omitted.

Lemma 7 *There exists a fixed partial gradient $\mathbf{v}_t \in \partial R(\mathbf{w}_t)$ such that the updates in Algorithm 2 are equivalent to the following gradient mappings,*

$$\begin{pmatrix} \mathbf{w}_t \\ \boldsymbol{\alpha}_t \end{pmatrix} = \Pi_{\mathbb{R}^d, \mathcal{Q}_{\boldsymbol{\alpha}}} \left(\begin{pmatrix} \mathbf{u}_{t-1} - \gamma(G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1}) + \lambda \mathbf{v}_t) \\ \boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\beta}_{t-1}) \end{pmatrix} \right)$$

and

$$\begin{pmatrix} \mathbf{u}_t \\ \boldsymbol{\beta}_t \end{pmatrix} = \Pi_{\mathbb{R}^d, \mathcal{Q}_{\boldsymbol{\alpha}}} \left(\begin{pmatrix} \mathbf{u}_{t-1} - \gamma(G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t) \\ \boldsymbol{\beta}_{t-1} + \gamma G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) \end{pmatrix} \right),$$

with initialization $\boldsymbol{\beta}_0 = \boldsymbol{\alpha}_0$.

Lemma 8 *For all $t = 1, 2, \dots$, and any $\mathbf{w} \in \mathbb{R}^d, \boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}$, we have*

$$\begin{aligned}
 & \gamma \begin{pmatrix} G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) + \lambda \mathbf{v}_t \\ -G_{\boldsymbol{\alpha}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) \end{pmatrix}^\top \begin{pmatrix} \mathbf{w}_t - \mathbf{w} \\ \boldsymbol{\alpha}_t - \boldsymbol{\alpha} \end{pmatrix} \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2 \\
 & \quad + \gamma^2 \|G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) - G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1})\|_2^2 - \frac{1}{2} \|\boldsymbol{\alpha}_t - \boldsymbol{\beta}_{t-1}\|_2^2.
 \end{aligned}$$

Proof [of Theorem 4 for Algorithm 2] Similar to proof of Theorem 4 for Algorithm 1, we have

$$\begin{aligned}
 & \gamma (F(\mathbf{w}_t, \boldsymbol{\alpha}_t) - F(\mathbf{w}, \boldsymbol{\alpha}_t) + F(\mathbf{w}_t, \boldsymbol{\alpha}) - F(\mathbf{w}_t, \boldsymbol{\alpha}_t)) \\
 & \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2 + \gamma^2 \|G_{\mathbf{w}}(\mathbf{w}_t, \boldsymbol{\alpha}_t) - G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\beta}_{t-1})\|_2^2 - \frac{1}{2} \|\boldsymbol{\alpha}_t - \boldsymbol{\beta}_{t-1}\|_2^2 \\
 & \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2 + \gamma^2 c \|\boldsymbol{\alpha}_t - \boldsymbol{\beta}_{t-1}\|_2^2 - \frac{1}{2} \|\boldsymbol{\alpha}_t - \boldsymbol{\beta}_{t-1}\|_2^2 \\
 & \leq \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_{t-1} \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_{t-1} \end{pmatrix} \right\|_2^2 - \frac{1}{2} \left\| \begin{pmatrix} \mathbf{w} - \mathbf{u}_t \\ \boldsymbol{\alpha} - \boldsymbol{\beta}_t \end{pmatrix} \right\|_2^2,
 \end{aligned}$$

where the last step follows the inequality (7) in Lemma 3 and the fact $\gamma = \sqrt{1/(2c)}$. By adding the inequalities of all iterations, we have

$$\sum_{t=1}^T (F(\mathbf{w}_t, \boldsymbol{\alpha}) - F(\mathbf{w}, \boldsymbol{\alpha}_t)) \leq \frac{\|\mathbf{w}\|_2^2 + \|\boldsymbol{\alpha}\|_2^2}{\sqrt{(2/c)} T}. \quad (16)$$

We complete the proof by using the definitions of $\widehat{\mathbf{w}}_T, \widehat{\boldsymbol{\alpha}}_T$, and the convexity-concavity of $F(\mathbf{w}, \boldsymbol{\alpha})$ with respect to \mathbf{w} and $\boldsymbol{\alpha}$, respectively. \blacksquare

Finally, we state the convergence bound for the objective $\mathcal{L}(\mathbf{w})$ in (2) in the following corollary.

Corrolary 9 Let \mathbf{w}^* be the optimal solution to (2), bounded by $\|\mathbf{w}^*\|_2^2 \leq D_1$, and $\|\boldsymbol{\alpha}\|_2^2 \leq D_2, \forall \boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}$. By running Algorithm 1 or 2 with T iterations, we have

$$\mathcal{L}(\widehat{\mathbf{w}}_T) - \mathcal{L}(\mathbf{w}^*) \leq \frac{D_1 + D_2}{\sqrt{(2/c)} T}.$$

Proof Let $\mathbf{w} = \mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} \mathcal{L}(\mathbf{w})$ and $\boldsymbol{\alpha}^* = \arg \max_{\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}} F(\widehat{\mathbf{w}}_T, \boldsymbol{\alpha})$ in Theorem 4, then we have

$$\max_{\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}} F(\widehat{\mathbf{w}}_T, \boldsymbol{\alpha}) - F(\mathbf{w}^*, \widehat{\boldsymbol{\alpha}}_T) \leq \frac{\|\mathbf{w}^*\|_2^2 + \|\boldsymbol{\alpha}^*\|_2^2}{\sqrt{(2/c)} T},$$

Since $\mathcal{L}(\mathbf{w}) = \max_{\boldsymbol{\alpha} \in \mathcal{Q}_{\boldsymbol{\alpha}}} F(\mathbf{w}, \boldsymbol{\alpha}) \geq F(\mathbf{w}, \widehat{\boldsymbol{\alpha}}_T)$, then we have

$$\mathcal{L}(\widehat{\mathbf{w}}_T) - \mathcal{L}(\mathbf{w}^*) \leq \frac{D_1 + D_2}{\sqrt{(2/c)} T}.$$

\blacksquare

Comparison with existing algorithms We compare the proposed algorithm to the Pegasos algorithm (Shalev-Shwartz et al., 2007) for minimizing the ℓ_2^2 regularized hinge loss. Although in this case both algorithms achieve a convergence rate of $O(1/T)$, their dependence on the regularization parameter λ is very different. In particular, the convergence rate of the proposed algorithm is $O((1 + n\lambda)/[\sqrt{2n\lambda T}])$ by noting that $\|\mathbf{w}^*\|_2^2 = O(1/\lambda)$, $\|\boldsymbol{\alpha}^*\|_2^2 \leq \|\boldsymbol{\alpha}^*\|_1 \leq n$, and $c = 1/n$, while the Pegasos algorithm has a convergence rate of $O(1/[\lambda T])$, where we ignore the logarithmic term $\ln(T)$. According to the common assumption of learning theory (Wu and Zhou, 2005; Smale and Zhou, 2003), the optimal λ is $O(n^{-1/(\tau+1)})$ if the probability measure can be approximated by the closure of RKHS \mathcal{H}_κ with exponent $0 < \tau \leq 1$. As a result, the convergence rate of the proposed algorithm is $O(\sqrt{n}/T)$ while the convergence rate of Pegasos is $O(n^{1/(1+\tau)}/T)$. Since $\tau \in (0, 1]$, the proposed algorithm is more efficient than the Pegasos algorithm, particularly when the number of training examples n is very large. This is verified by our empirical studies in section 5.6 (see Figure 8).

4.4 Implementation

We briefly discuss how to efficiently solve the optimization problems for updating the primal and dual variables in Algorithms 1 and 2. Both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are updated by a gradient mapping that requires computing the projection into the domain \mathcal{Q}_α . When \mathcal{Q}_α is only consisted of box constraints (e.g. hinge loss, absolute loss, and ϵ -insentive loss), the projection $\Pi_{\mathcal{Q}_\alpha}[\hat{\alpha}]$ can be computed by thresholding. When \mathcal{Q}_α is comprised of both box constraints and a linear constraint (e.g. generalized hinge loss), the following lemma gives an efficient algorithm for computing $\Pi_{\mathcal{Q}_\alpha}[\hat{\alpha}]$.

Lemma 10 *For $\mathcal{Q}_\alpha = \{\boldsymbol{\alpha} : \boldsymbol{\alpha} \in [0, s]^n, \boldsymbol{\alpha}^\top \mathbf{v} \leq \rho\}$, the optimal solution $\boldsymbol{\alpha}^*$ to projection $\Pi_{\mathcal{Q}_\alpha}[\hat{\alpha}]$ is computed by*

$$\alpha_i^* = [\hat{\alpha}_i - \eta v_i]_{[0, s]}, \forall i \in [n],$$

where $\eta = 0$ if $\sum_i [\hat{\alpha}_i]_{[0, s]} v_i \leq \rho$ and otherwise is the solution to the following equation

$$\sum_i [\hat{\alpha}_i - \eta v_i]_{[0, s]} v_i - \rho = 0. \quad (17)$$

Since $\sum_i [\hat{\alpha}_i - \eta v_i]_{[0, s]} v_i - \rho$ is monotonically decreasing in η , we can solve η in (17) by a bi-section search.

For the composite gradient mapping for $\mathbf{w} \in \mathcal{Q}_\mathbf{w} = \mathbb{R}^d$, there is a closed form solution for simple regularizers (e.g., ℓ_1, ℓ_2) and decomposable regularizers (e.g., $\ell_{1,2}$). Efficient algorithms are available for composite gradient mapping when the regularizer is the ℓ_∞ and $\ell_{1,\infty}$, or trace norm. More details can be found in (Duchi and Singer, 2009; Ji and Ye, 2009). Here we present an efficient solution to a general regularizer $V(\|\mathbf{w}\|)$, where $\|\mathbf{w}\|$ is either a simple regularizer (e.g., ℓ_1, ℓ_2 , and ℓ_∞) or a decomposable regularizer (e.g., $\ell_{1,2}$ and $\ell_{1,\infty}$), and $V(z)$ is convex and monotonically increasing for $z \geq 0$. An example is $V(\mathbf{w}_1, \dots, \mathbf{w}_K) = (\sum_k \|\mathbf{w}_k\|_2)^2$, where $\mathbf{w}_1, \dots, \mathbf{w}_K$ forms a partition of \mathbf{w} .

Lemma 11 *Let $V_*(\eta)$ be the convex conjugate of $V(z)$, i.e. $V(z) = \max_{\eta} \eta z - V_*(\eta)$. Then the solution to the composite mapping*

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} \frac{1}{2} \|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2 + \lambda V(\|\mathbf{w}\|),$$

can be computed by

$$\mathbf{w}^* = \arg \min_{\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}} \frac{1}{2} \|\mathbf{w} - \widehat{\mathbf{w}}\|_2^2 + \lambda \eta \|\mathbf{w}\|,$$

where η satisfies $\|\mathbf{w}^*\| - V'_*(\eta) = 0$. Since both $\|\mathbf{w}^*\|$ and $-V'_*(\eta)$ are non-increasing functions in η , we can efficiently compute η by a bi-section search.

4.5 Extensions and Discussion

Nonlinear model For a nonlinear model, the min-max formulation becomes

$$\min_{g \in \mathcal{H}_{\kappa}} \max_{\alpha \in \mathcal{Q}_{\alpha}} L(g, \alpha; \mathbf{X}, \mathbf{y}) + \lambda R(g),$$

where \mathcal{H}_{κ} is a Reproducing Kernel Hilbert Space (RKHS) endowed with a kernel function $\kappa(\cdot, \cdot)$. Algorithm 1 can be applied to obtain the nonlinear model by changing the primal variable to g . For example, step 5 in Algorithm 1 is modified to the following composite gradient mapping

$$g_t = \arg \min_{g \in \mathcal{H}_{\kappa}} \frac{1}{2} \|g - \hat{g}_{t-1}\|_{\mathcal{H}_{\kappa}}^2 + \gamma \lambda R(g),$$

where

$$\hat{g}_{t-1} = (g_{t-1} - \gamma \nabla_g L(g_{t-1}, \alpha_t; \mathbf{X}, \mathbf{y})).$$

Similar changes can be made to Algorithm 2 for the extension to the nonlinear model.

Incorporating the bias term Algorithms 1 or 2 can be modified to incorporate the bias term b by viewing b as an additional primal variable. For example, in Algorithm 1, at each iteration, we can update b by one prox mapping $b_t = \Pi_{\mathcal{Q}_b}[b_{t-1} + \gamma \nabla_b L(\mathbf{w}_{t-1}, b_{t-1}, \alpha_t; \mathbf{X}, \mathbf{y})]$, where \mathcal{Q}_b is an appropriate domain for b . The step size γ is changed to $\gamma = \sqrt{(1/c)}/2$ when the bias term is considered, and the convergence rate is amplified by a constant $\sqrt{2}$.

Domain constraint on primal variable Now we discuss how to generalize the convergence analysis to the case when a convex domain $\mathcal{Q}_{\mathbf{w}}$ is imposed on \mathbf{w} . We introduce $\widehat{R}(\mathbf{w}) = \lambda R(\mathbf{w}) + Q(\mathbf{w})$, where $Q(\mathbf{w})$ is an indicator function for $\mathbf{w} \in \mathcal{Q}_{\mathbf{w}}$, i.e.

$$Q(\mathbf{w}) = \begin{cases} 0 & \mathbf{w} \in \mathcal{Q}_{\mathbf{w}} \\ +\infty & \text{otherwise} \end{cases}.$$

Then we can write the domain constrained composite gradient mapping in step 5 of Algorithm 1 or step 4 of Algorithm 2 into a domain free composite gradient mapping as the following:

$$\begin{aligned} \mathbf{w}_t &= \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w} - (\mathbf{w}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{w}_{t-1}, \alpha_t))\|_2^2 + \gamma \widehat{R}(\mathbf{w}), \\ \mathbf{w}_t &= \arg \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{2} \|\mathbf{w} - (\mathbf{u}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{u}_{t-1}, \alpha_{t-1}))\|_2^2 + \gamma \widehat{R}(\mathbf{w}). \end{aligned}$$

Then we have an equivalent gradient mapping,

$$\begin{aligned}\mathbf{w}_t &= \mathbf{w}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{w}_{t-1}, \boldsymbol{\alpha}_t) - \gamma \partial \hat{R}(\mathbf{w}_t), \\ \mathbf{w}_t &= \mathbf{u}_{t-1} - \gamma G_{\mathbf{w}}(\mathbf{u}_{t-1}, \boldsymbol{\alpha}_{t-1}) - \gamma \partial \hat{R}(\mathbf{w}_t).\end{aligned}$$

Then Lemmas 5 and 6, and Lemmas 7 and 8 all hold as long as we replace $\lambda \mathbf{v}_t$ with $\hat{\mathbf{v}}_t \in \partial \hat{R}(\mathbf{w}_t)$. Finally in proving Theorems 4 we can absorb $Q(\mathbf{w})$ in $L(\mathbf{w}, \boldsymbol{\alpha}) + \hat{R}(\mathbf{w})$ into the domain constraint.

Additional constraints on dual variables One advantage of the proposed primal dual prox method is that it provides a convenient way to handle additional constraints on the dual variables α . Several studies introduce additional constraints on the dual variables. In (Dekel and Singer, 2006), the authors address a budget SVM problem by introducing a $1 - \infty$ interpolation norm on the empirical hinge loss, leading to a sparsity constraint $\|\boldsymbol{\alpha}\|_1 \leq m$ on the dual variables, where m is the target number of support vectors. The corresponding optimization problem is given by

$$\min_{\mathbf{w} \in \mathbb{R}^d} \max_{\boldsymbol{\alpha} \in [0,1]^n, \|\boldsymbol{\alpha}\|_1 \leq m} \frac{1}{n} \sum_{i=1}^n \alpha_i (1 - y_i \mathbf{w}^\top \mathbf{x}_i) + \lambda R(\mathbf{w}). \quad (18)$$

In (Huang et al., 2010), a similar idea is applied to learn a distance metric from noisy training examples. We can directly apply Algorithms 1 or 2 to (18) with $\mathcal{Q}_{\boldsymbol{\alpha}}$ given by $\mathcal{Q}_{\boldsymbol{\alpha}} = \{\boldsymbol{\alpha} : \boldsymbol{\alpha} \in [0,1]^n, \|\boldsymbol{\alpha}\|_1 \leq m\}$. The prox mapping to this domain can be efficiently computed by Lemma 10. It is straightforward to show that the convergence rate is $[D_1 + m]/[\sqrt{2nT}]$ in this case.

5. Experiments

In this section we present empirical studies to verify the efficiency of the proposed algorithm. We organize our experiments as follows.

- In subsections 5.1, 5.2, and 5.3 we compare the proposed algorithm to the state-of-the-art first order methods that directly update the primal variable at each iteration. We apply all the algorithms to three different tasks with different non-smooth loss functions and regularizers. The baseline first order methods used in this study include the gradient descent algorithm (**gd**), the forward and backward splitting algorithm (**fobos**) (Duchi and Singer, 2009), the regularized dual averaging algorithm (**rda**) (Xiao, 2009), the accelerated gradient descent algorithm (**agd**) (Chen et al., 2009). Since the proposed method is a non-stochastic method, we compare it to the non-stochastic variant of **gd**, **fobos**, and **rda**. Note that **gd**, **fobos**, **rda**, and **agd** share the same convergence rate of $O(1/\sqrt{T})$ for non-smooth problems.
- In subsection 5.4, our algorithm is compared to the state-of-the-art primal dual gradient method (Nesterov, 2005b), which employs an excessive gap technique for non-smooth optimization, updates both the primal and dual variables at each iteration, and has a convergence rate of $O(1/T)$.

- In subsection 5.5, we test the proposed algorithm for optimizing problem in (18) with a sparsity constraint on the dual variables.
- In subsection 5.6, we compare the two variants of the proposed method on a data set when $n \gg d$, and compare Pdprox to Pegasos algorithm.

All the algorithms are implemented in Matlab and run on a 2.4GHZ machine. Since the performance of the baseline algorithms **gd**, **fobos** and **rda** depends heavily on the initial value of the stepsize, we generate 21 values for the initial stepsize by scaling their theoretically optimal values with factors $2^{[-10:1:10]}$, and report the best convergence among the 21 possible values. The stepsize of **agd** is changed adaptively in the optimization process, and we just give it an appropriate initial step size. Since in the first four subsections we focus on comparison to baselines, we use the Pdprox-dual variant of the proposed Pdprox algorithm. Finally, all algorithms are initialized with a solution of all zeros.

5.1 Group lasso regularizer for Grouped Feature Selection

In this experiment we use the group lasso for regularization, i.e., $R(\mathbf{w}) = \sum_g \sqrt{d_g} \|\mathbf{w}_g\|_2$, where \mathbf{w}_g corresponds to the g th group variables and d_g is the number of variables in group g . To apply Nesterov’s method, we can write $R(\mathbf{w}) = \max_{\|\mathbf{u}_g\|_2 \leq 1} \sum_g \sqrt{d_g} \mathbf{w}_g^\top \mathbf{u}_g$. We use the MEMset Donar dataset (Yeo and Burge, 2003) as the testbed. This dataset was originally used for splice site detection. It is divided into a training set and a test set: the training set consists of 8,415 true and 179,438 false donor sites, and the testing set has 4,208 true and 89,717 false donor sites. Each example in this dataset was originally described by a sequence of {A, C, G, T} of length 7. We follow (Yang et al., 2010) and generate group features with up to three-way interactions between the 7 positions, leading to 2,604 attributes in 63 groups. We normalize the length of each example to 1. Following the experimental setup in (Yang et al., 2010), we construct a balanced training dataset consisted of all 8,415 true and 8,415 false donor sites that are randomly sampled from all 179,438 false sites.

Two non-smooth loss functions are examined in this experiment: hinge loss and absolute loss. Figure 2 plots the values of the objective function vs. running time (second), using two different values of regularization parameter, i.e., $\lambda = 10^{-3}, 10^{-5}$ to produce different levels of sparsity. We observe that (i) the proposed algorithm **Pdprox** clearly outperforms all the baseline algorithms in all the cases; (ii) for the absolute loss, which has a sharp curvature change at zero compared to hinge loss, the baseline algorithms of **gd**, **fobos**, **rda**, **agd**, especially of **agd** that is originally designed for smooth loss functions, deteriorate significantly compared to the proposed algorithm **Pdprox**. Finally, we observe that for the hinge loss and $\lambda = 10^{-3}$, the classification performance of the proposed algorithm on the testing dataset is 0.6565, measured by maximum correlation coefficient (Yeo and Burge, 2003). This is almost identical to the best result reported in (Yang et al., 2010) (i.e., 0.6520).

5.2 $\ell_{1,\infty}$ regularization for Multi-Task Learning

In this experiment, we perform multi-task regression with $\ell_{1,\infty}$ regularizer (Chen et al., 2009). Let $\mathbf{W} = (\mathbf{w}_1, \dots, \mathbf{w}_k) \in \mathbb{R}^{d \times k}$ denote the k linear hypotheses for regression. The

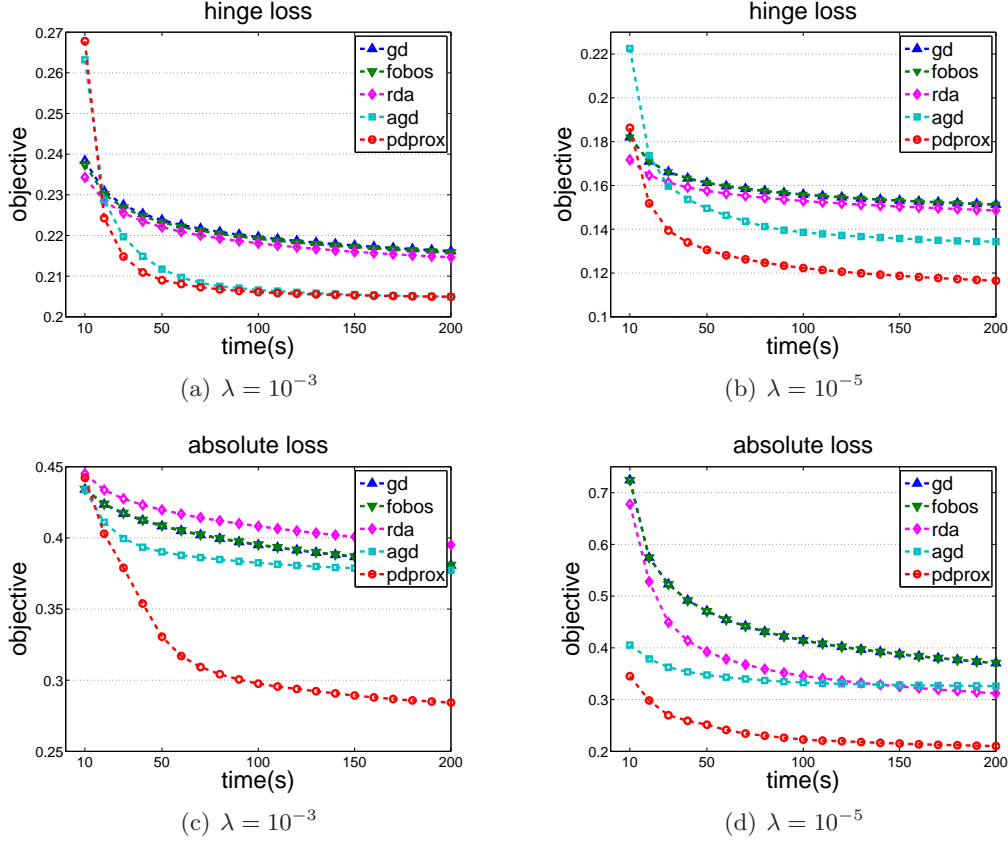


Figure 2: Comparison of convergence speed for hinge loss ((a),(b)) and absolute loss ((c),(d)) with group lasso regularizer. Note that for better visualization we plot the objective starting from 10 seconds in all figures. The objective of all algorithms at 0 second is 1.

$\ell_{1,\infty}$ regularizer is given by $R(\mathbf{W}) = \sum_{j=1}^d \|\mathbf{w}^j\|_\infty$, where \mathbf{w}^j is the j th row of \mathbf{W} . To apply Nesterov’s method, we rewrite the $\ell_{1,\infty}$ regularizer as $R(\mathbf{W}) = \max_{\|\mathbf{u}_j\|_1 \leq 1} \sum_{j=1}^d \mathbf{u}_j^\top \mathbf{w}^j$. We use the School data set (Argyriou et al., 2008), a common dataset for multi-task learning. This data set contains the examination scores of 15,362 students from 139 secondary schools corresponding to 139 tasks, one for each school. Each student in this dataset is described by 27 attributes. We follow the setup in (Argyriou et al., 2008), and generate a training data set with 75% of the examples from each school and a testing data set with the remaining examples. We test the algorithms using both the absolute loss and the ϵ -insensitive loss with $\epsilon = 0.01$. The initial stepsize for **gd**, **fobos**, and **rda** are tuned similarly as that for the experiment of group lasso. We plot the objective versus the running time in Figure 3, from which we observe the similar results in the group feature selection task, i.e. (i) the proposed **Pdprox** algorithm outperforms the baseline algorithms, (ii) the baseline algorithm of **agd** becomes even worse for ϵ -insensitive loss than for absolute loss. Finally, we observe that the regression performance measured by root mean square error (RMSE) on

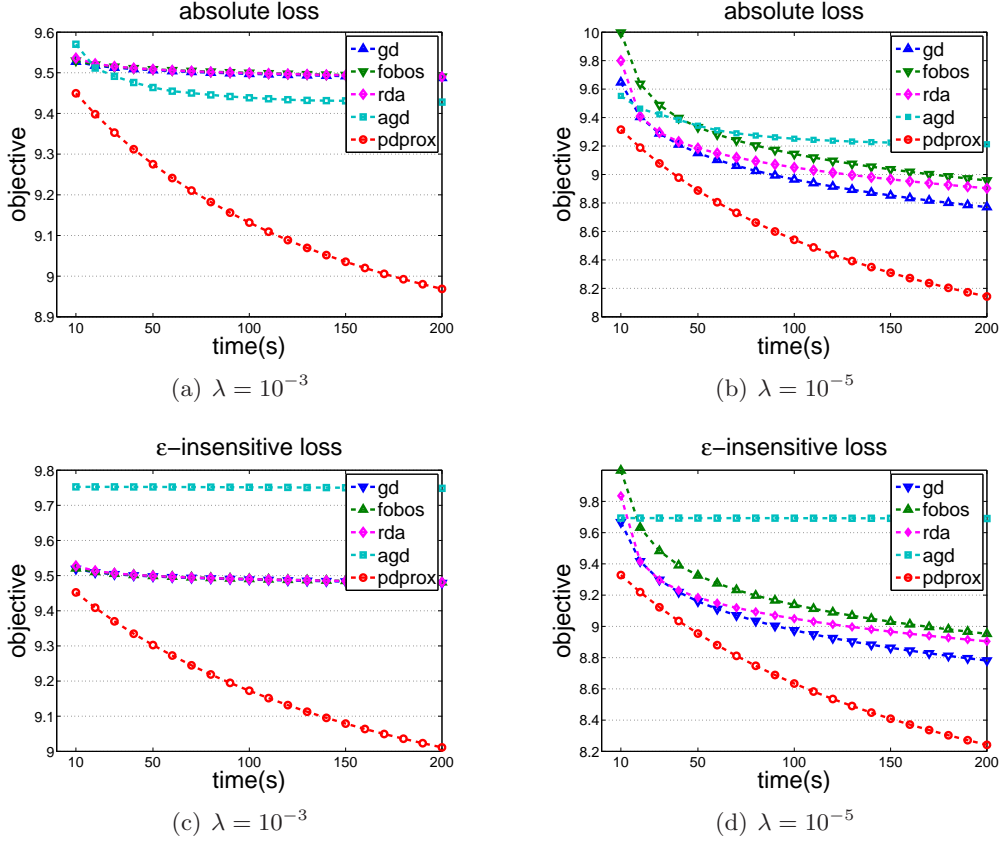


Figure 3: Comparison of convergence speed for absolute loss ((a),(b)) and ϵ -insensitive loss ((c),(d)) with $\ell_{1,\infty}$ regularizer. Note that for better visualization we plot the objective starting from 10 seconds in all figures. The objective of all algorithms at 0 second is 20.52.

the testing data set for absolute loss and ϵ -insensitive loss is 10.34 (optimized by **Pdprox**), comparable to the performance reported in (Chen et al., 2009).

5.3 Trace norm regularization for Max-Margin Matrix Factorization/ Matrix Completion

In this experiment, we evaluate the proposed method using trace norm regularization, a regularizer often used in max-margin matrix factorization and matrix completion, where the goal is to recover a full matrix \mathbf{X} from partially observed matrix \mathbf{Y} . The objective is composed of a loss function measuring the difference between \mathbf{X} and \mathbf{Y} on the observed entries and a trace norm regularizer on \mathbf{X} , assuming that \mathbf{X} is low rank. Hinge loss function is used in max-margin matrix factorization (Rennie and Srebro, 2005; Srebro et al., 2005), and absolute loss is used instead of square loss in matrix completion. We test on 100K Movie-

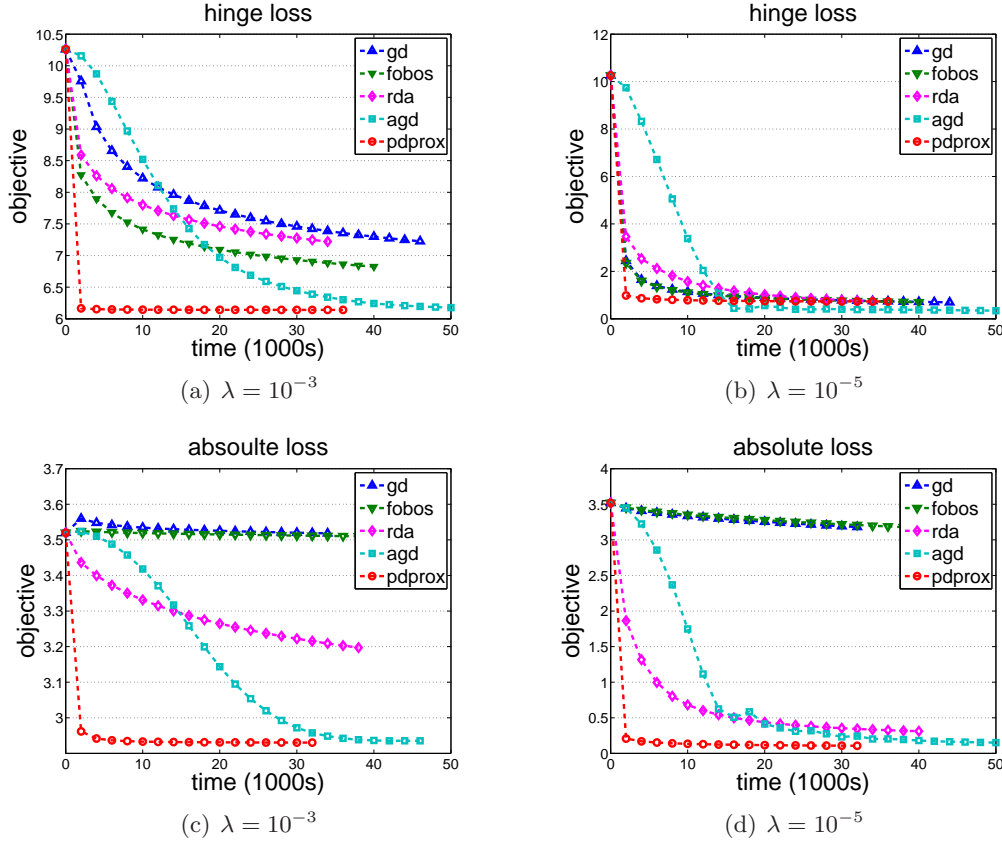


Figure 4: Comparison of convergence speed for (a,b): max-margin matrix factorization with hinge loss and trace norm regularizer, and (c,d): matrix completion with absolute loss and trace norm regularizer.

Lens data set ¹ that contains 1 million ratings from 943 users on 1682 movies. Since there are five distinct ratings that can be assigned to each movie, we follow (Rennie and Srebro, 2005; Srebro et al., 2005) by introducing four thresholds $\theta_{1,2,3,4}$ to measure the hinge loss between the predicted value X_{ij} and the ground truth Y_{ij} . Because our goal is to demonstrate the efficiency of the proposed algorithm for non-smooth optimization, therefore we simply set $\theta_{1,2,3,4} = (0, 3, 6, 9)$. Note that we did not compare to the optimization algorithm in (Rennie and Srebro, 2005) since it cast the problem into a non-convex problem by using explicit factorization of $\mathbf{X} = \mathbf{UV}^\top$, which suffers a local minimum, and the optimization algorithm in (Srebro et al., 2005) since it formulated the problem into a SDP problem, which suffers from a high computational cost. To apply Nesterov's method, we write $\|\mathbf{X}\|_1 = \max_{\|\mathbf{A}\| \leq 1} \text{tr}(\mathbf{A}^\top \mathbf{X})$, and at each iteration we need to solve a maximization problem $\max_{\|\mathbf{A}\| \leq 1} \lambda \text{tr}(\mathbf{A}^\top \mathbf{X}) - \frac{\mu}{2} \|\mathbf{A}\|_F^2$, where $\|\mathbf{A}\|$ is the spectral norm on \mathbf{A} . The solution of this optimization is obtained by performing SVD decomposition of \mathbf{X} and thresholding the singular values appropriately. Since MoiveLens data set is much larger than the data

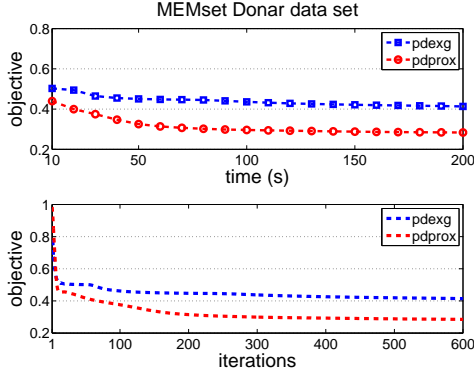
1. <http://www.cs.umn.edu/Research/GroupLens/>

sets used in last two subsections, in this experiment, we (i) run all the algorithms 1000 iterations and plot the objective versus time; (ii) enlarge the range of tuning parameters to $2^{[-15:1:15]}$. The results are shown in Figure 4, from which we observe that (i) Pdprox can quickly reduce the objective in a small amount of time, e.g., for absolute loss when setting $\lambda = 10^{-3}$ in order to obtain a solution with an accuracy of 10^{-3} , **Pdprox** needs 10^3 second, while **agd** needs 3.2×10^4 seconds; (ii) for absolute loss no matter how we tune the stepsizes for each baseline algorithm, Pdprox performs the best; and (iii) for hinge loss when $\lambda = 10^{-5}$, by tuning the stepsizes of baseline algorithms, **gd**, **fobos**, and **rda** can achieve comparable performance to Pdprox. We note that although **agd** can achieve smaller objective value than Pdprox at the end of 1000 iterations, however, the objective value is reduced slowly.

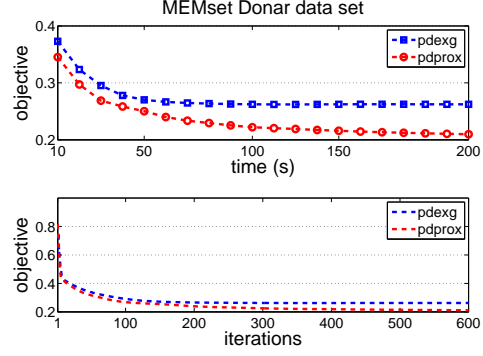
5.4 Comparison: Pdprox vs Primal-Dual method with excessive gap technique

In this section, we compare the proposed primal dual prox method to Nesterov’s primal dual method (Nesterov, 2005b), which is an improvement of his algorithm in (Nesterov, 2005a). The algorithm in (Nesterov, 2005a) for non-smooth optimization suffers a problem of setting the value of smoothing parameter that requires the number of iterations to be fixed in advance. Nesterov (2005b) addresses the problem by exploring an excessive gap technique and updating both the primal and dual variables, which is similar to the proposed Pdprox method. We refer to this baseline as **Pdexg**. We run both algorithms on the three tasks as in subsections 5.1, 5.2, and 5.3, i.e., group feature selection with hinge loss and group lasso regularizer on MEMset Donar data set, multi-task learning with ϵ -insensitive loss and $\ell_{1,\infty}$ regularizer on School data set, and matrix completion with absolute loss and trace norm regularizer on 100K MoiveLens data set. To implement the primal dual method with excessive gap technique, we need to intentionally add a domain on the optimal primal variable, which can be derived from the formulation. For example, in group feature selection problem whose objective is $1/n \sum_{i=1}^n \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda \sum_g \sqrt{d_g} \|\mathbf{w}_g\|_2$, we can derive that the optimal primal variable \mathbf{w}^* lies in $\|\mathbf{w}\|_2 \leq \sum_g \|\mathbf{w}_g\|_2 \leq \frac{1}{\lambda \sqrt{d_{\min}}}$, where $d_{\min} = \min_g d_g$. Similar techniques are applied to multi-task learning and matrix completion.

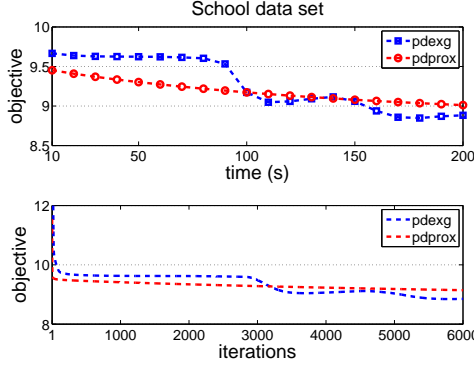
The performance of the two algorithms on the three tasks is shown in Figure 5. Since both algorithms are in the same category, i.e. updating both primal and dual variables at each iteration and having a convergence rate in the order of $O(1/T)$, we also plot the objective versus the number of iterations in the bottom panels of each subfigure in Figure 5. The results show that the proposed Pdprox method converges faster than Pdexg on MEMset Donar data set for group feature selection with hinge loss and group lasso regularizer, and on 100K MovieLens data set for matrix completion with absolute loss and trace norm regularizer. However, Pdexg performs better on School data set for multi-task learning with ϵ -insensitive loss and $\ell_{1,\infty}$ regularizer. One interesting phenomenon we can observe from Figure 5 is that for larger values of λ (e.g., 10^{-3}), the improvement of Pdprox over Pdexg is also larger. The reason is because that the proposed Pdprox captures the sparsity of primal variable at each iteration. This does not hold for Pdexg because it casts the non-smooth regularizer into a dual form and consequently does not explore the sparsity of the primal variable at each iteration. Therefore the larger of λ , the sparser of the primal variable at each iteration in Pdprox that yields to larger improvement over Pdexg. For the



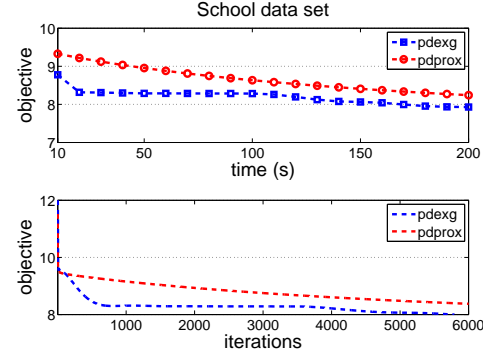
(a) Group feature selection: hinge loss and group lasso regularizer with $\lambda = 10^{-3}$.



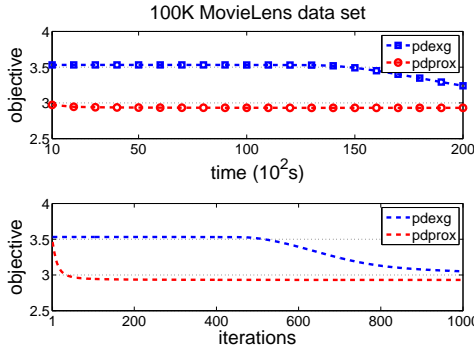
(b) Group feature selection: hinge loss and group lasso regularizer with $\lambda = 10^{-5}$.



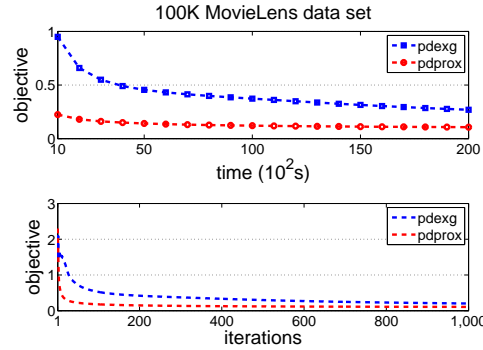
(c) Multi-task learning: ϵ -insensitive loss and $\ell_{1,\infty}$ regularizer with $\lambda = 10^{-3}$.



(d) Multi-task learning: ϵ -insensitive loss and $\ell_{1,\infty}$ regularizer with $\lambda = 10^{-5}$.



(e) Matrix completion: absolute loss and trace norm regularizer with $\lambda = 10^{-3}$.



(f) Matrix completion: absolute loss and trace norm regularizer with $\lambda = 10^{-5}$.

Figure 5: Pdprox vs Primal-Dual method with excessive gap technique.

example of group feature selection task with hinge loss and group lasso regularizer, when setting $\lambda = 10^{-3}$, the sparsity of the primal variable (i.e., the proportion of the number of group features with zero norm) in Pdprox averaged over all iterations is 0.7886. However, by

Table 1: Running time (forth column) and classification accuracy (fifth column) of Pdprox for (18) and of Liblinear on noisily labeled training data, where noise is added to labels by random flipping with a probability 0.2. We fix $\lambda = 1/n$ or $C = 1$ in Liblinear. In the second column, we report the number of training examples (n), the number of attributes (d), and also the accuracy by training Liblinear on the original data and evaluating it on the testing data.

Data Set	$(n, d)/\text{ACC}$	Alg.	Running time	ACC
a9a	(32561, 123)	Pdprox($m=200$)	0.82s(0.013)	0.8344(0.0005)
	0.8501	Liblinear	1.15s(0.5688)	0.7890(0.0044)
rcv1	(20242, 47236)	Pdprox($m=200$)	1.57s(0.2328)	0.9405(0.0022)
	0.9654	Liblinear	3.30s(0.7399)	0.9366(0.0015)
covtype	(571012, 54)	Pdprox($m=4000$)	48s(3.34)	0.7358(10^{-4})
	0.7580	Liblinear	37s(0.64)	0.6866(10^{-5})

reducing λ to 10^{-5} the average sparsity of the primal variable in Pdprox is reduced to 0. In both settings the average sparsity of the primal variable in Pdexg is 0. The same argument also explains why Pdprox does not perform as well as Pdexg on School data set when setting $\lambda = 10^{-5}$, since in this case the primal variables in both algorithms are not sparse. When setting $\lambda = 10^{-3}$, the average sparsity (i.e., the proportion of the number of features with zero norm across all tasks) of the primal variable in Pdprox and Pdexg is 0.3766 and 0, respectively. Finally, we also observe similar performance of the two algorithms on the three tasks with other loss functions including absolute loss for group feature selection, absolute loss for multi-task learning, and hinge loss for max-margin matrix factorization.

5.5 Sparsity constraint on the dual variables

In this subsection, we examine empirically the proposed algorithm for optimizing the problem in equation (18), in which a sparsity constraint is introduced for the dual variables. We test the algorithm on three large data sets from the UCI repository, namely, a9a, rcv1(binary) and covtype². In the experiments we use ℓ_2^2 regularizer and fix $\lambda = 1/n$. First, we run the proposed algorithm 100 seconds on the three data sets with different values of $m = 100, 200, 400$ and plot the objective versus the number of iterations. The results are shown in Figure 6, which verify our theoretical convergence bound (i.e., $O([D+m]/[\sqrt{2n\lambda}])$) of the proposed algorithm for (18), i.e., the smaller the m , the faster the convergence.

Second, we demonstrate that the formulation in equation (18) with a sparsity constraint on the dual variables is useful in the case when labels are contaminated with noise. To generate the noise in labels, we randomly flip the labels with a probability 0.2. We run both the proposed algorithm for (18) and Liblinear³ on the training data with noise added to the labels. The stopping criterion for the proposed algorithm is when duality gap is less than 10^{-3} , and for Liblinear is when the maximal dual violation is less than 10^{-3} . The

2. <http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets>

3. <http://www.csie.ntu.edu.tw/~cjlin/liblinear>

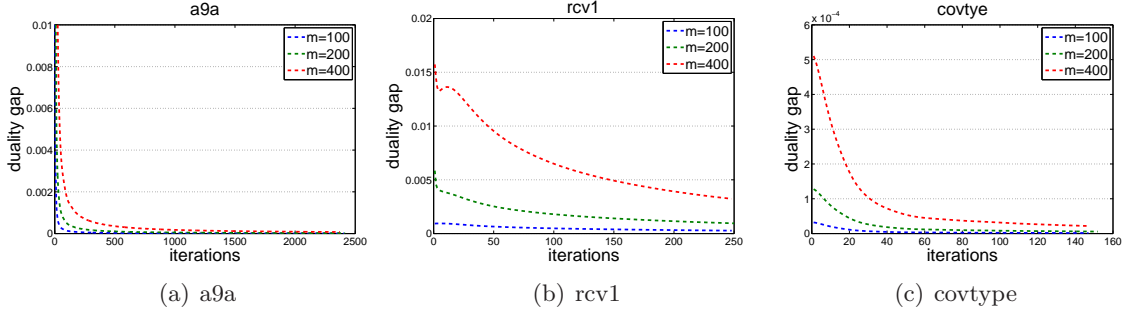
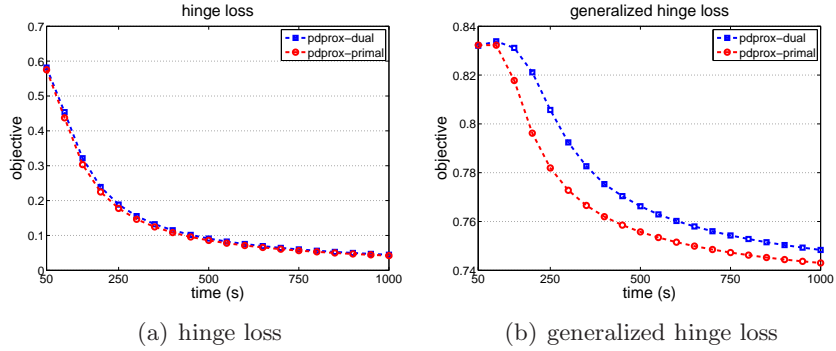

 Figure 6: Comparison of Convergence with varied m .


Figure 7: Comparison of Pdprox-dual vs. Pdprox-primal on covtype data set.

running time and accuracy on testing data averaged over 5 random trials are reported in Table 1, which demonstrate that in the presence of noise in labels, by adding a sparsity constraint on the dual variables, we are able to obtain better performance than Liblinear trained on the noisily labeled data. Furthermore the running time of Pdprox is comparable to, if not less than, that of Liblinear.

Finally, we note that choosing a small m in equation (18) is different from simply training a classifier with a small number of examples. For instance, for rcv1, we have run the experiment with 200 training examples, randomly selected from the entire data set. With the same stopping criterion, the testing performance is $0.8131(\pm 0.05)$, significantly lower than that of optimizing (18) with $m = 200$.

5.6 Comparisons: Pdprox-dual vs. Pdprox-primal and Pdprox vs. Pegasos

In this subsection, we verify (i) the Pdprox-primal is more efficient than Pdprox-dual when $n \gg d$, and (ii) the proposed Pdprox method converges faster than Pegasos when $\lambda = O(n^{-1/(1+\tau)})$, $\tau \in (0, 1]$. For a fair comparison, we use ℓ_2^2 norm for the regularizer since Pegasos is designed for ℓ_2^2 regularizer. We choose the covtype data set as testing bed, which contains $n = 581012$ examples, and each example is represented by a vector of $d = 54$ dimension.

Figure 7 shows the comparison of **Pdprox-dual** and **Pdprox-primal** on covtype data set for two loss functions, i.e. hinge loss and generalized hinge loss. The domains for the

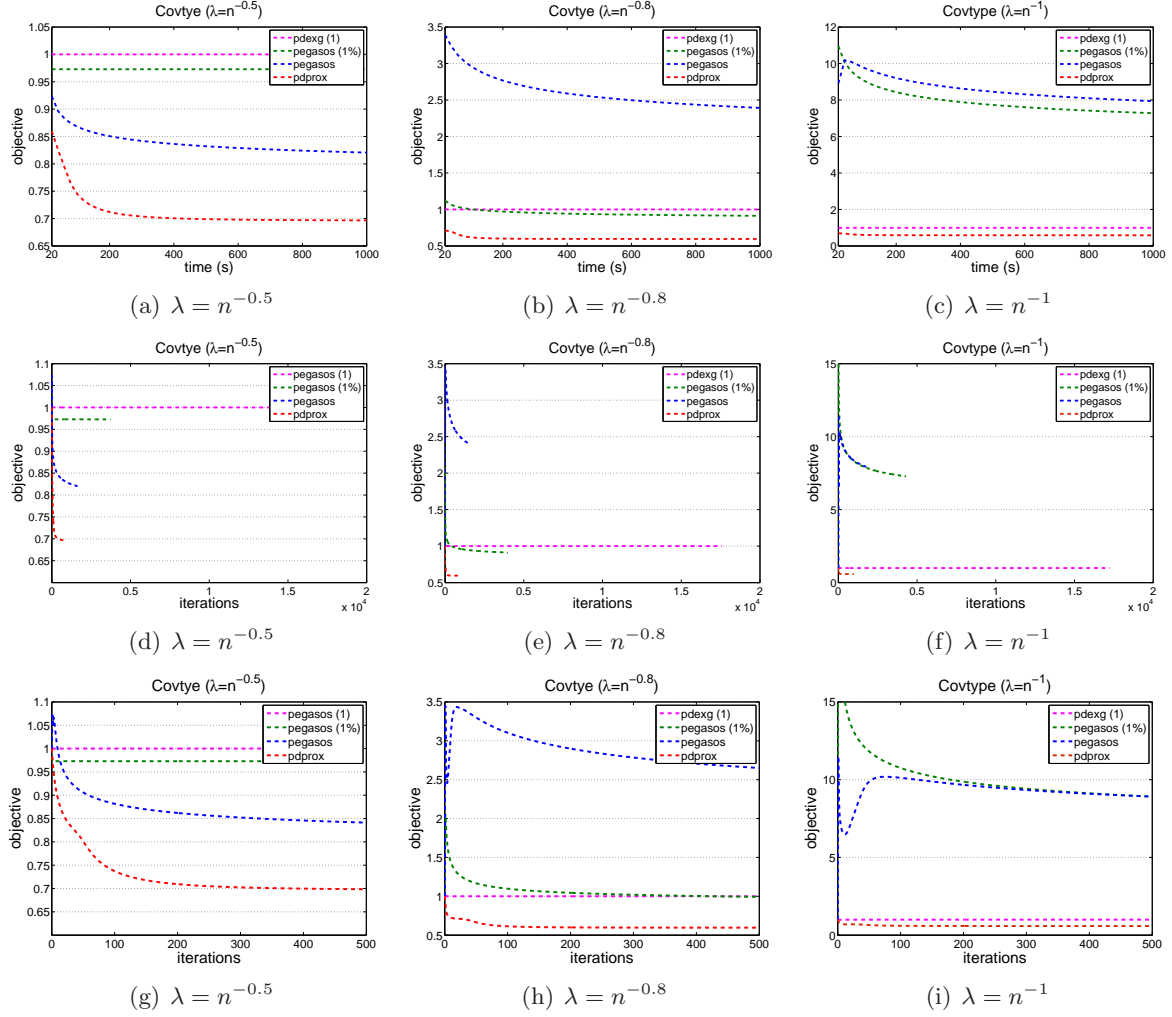


Figure 8: Comparison of convergence speed of Pdprox vs. Pegasos on covtype data set.

dual variables of these two loss functions are $\mathcal{Q}_\alpha = \{\alpha : \alpha \in [0, 1]^n\}$ and $\mathcal{Q}_\alpha = \{(\alpha^1, \alpha^2) \in [0, 1]^{n \times 2}, \alpha^1 + \alpha^2 \leq \mathbf{1}\}$, respectively. We can see that the domain for generalized hinge loss is complex than that for hinge loss, and we observe that Pdprox-primal is more efficient than Pdprox-dual for generalized hinge loss, since the former one maintains one copy of dual variables and only needs to compute one projection of the dual variables to the corresponding domain while the latter one maintains two copies of the dual variables and requires two projections. However, the two algorithms perform almost the same on simple domain of dual variables for hinge loss.

Figure 8 shows the comparison of **Pdprox** vs. **Pegasos** on Covtype data sets with three different levels of $\lambda = n^{-0.5}, n^{-0.8}, n^{-1}$. For Pdprox, we use the Pdprox-primal algorithm. Since Pegasos can perform stochastic updating, we also report the performance of Pegasos algorithm by sampling 1 example and 1% examples for computing the stochastic gradient. We run both algorithms 1000 seconds, report the objective versus time in the top panel

of Figure 8, and report the objective versus the number of iterations (achieved in 1000 seconds) in the middle panel. In the bottom panel, we also report the objective versus the first 500 iterations to better understand the behavior of algorithms at the beginning of iterations. The results show that (i) Pdprox converges faster than Pegasos when $\lambda = O(n^{-1/(1+\tau)})$, $\tau = 0, 0.25, 1$, which verifies our discussion in subsection 4.5; (ii) the objective of Pegasos always has a sharp jump at the beginning of iterations, which becomes severe when λ is reduced. This is because at the beginning, the gradient of the primal variable is not informative compared to that in Pdprox which uses the updated dual variables to compute the gradient. Finally, we observe that the performance of stochastic Pegasos with sampling 1% examples becomes better than the performance of batch Pegasos when λ is reduced to be less than $n^{-0.8}$, however, stochastic Pegasos performs consistently worse than Pdprox.

6. Conclusions

In this paper, we study non-smooth optimization in machine learning where both the loss function and the regularizer are non-smooth. We develop an efficient gradient based method for a family of non-smooth optimization problems in which the dual form of the loss function can be expressed as a bilinear function in primal and dual variables. We show that the proposed algorithm achieves a convergence rate of $O(1/T)$, significantly more efficient than the other existing first order methods for non-smooth optimization. Our empirical studies demonstrate the efficiency of the proposed algorithm in comparison to the state-of-the-art first order methods for non-smooth optimization problems, and the effectiveness of the proposed algorithm for optimizing the problem with a sparse constraint on the dual variables in the presence of noisy labels. In future, we plan to adapt the proposed algorithm for stochastic updating and for distributed computing environments.

References

- Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73:243–272, 2008.
- Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 9:1823–1840, 2008.
- Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- Yunpeng Cai, Yijun Sun, Yubo Cheng, Jian Li, and Steve Goodison. Fast implementation of l1 regularized learning algorithms using gradient descent methods. In *SDM*, pages 862–871, 2010.
- Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *CoRR*, abs/0805.4471, 2008.
- Xi Chen, Weike Pan, James T. Kwok, and Jaime G. Carbonell. Accelerated gradient method for multi-task sparse learning problem. In *ICDM*, pages 746–751, 2009.

- Ofer Dekel and Yoram Singer. Support vector machines on a budget. In *NIPS*, pages 345–352, 2006.
- John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *JMLR.*, 10:2899–2934, 2009.
- Glenn Fung and O. L. Mangasarian. A feature selection newton method for support vector machine classification. Technical report, Computational Optimization and Applications, 2002.
- Tilmann Gneiting. Quantiles as optimal point predictors. Technical report, Department of Statistics, University of Washington, 2008.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2008.
- Chenping Hou, Feiping Nie, Dongyun Yi, and Yi Wu. Feature selection via joint embedding learning and sparse regression. In *IJCAI*, pages 1324–1329, 2011.
- Cho-Jui Hsieh, Kai-Wei Chang, Chih-Jen Lin, S. Sathya Keerthi, and S. Sundararajan. A dual coordinate descent method for large-scale linear svm. In *ICML*, pages 408–415, 2008.
- Chonghai Hu, James Kwok, and Wei-ke Pan. Accelerated gradient methods for stochastic optimization and online learning. In *NIPS*, pages 781–789, 2009.
- Kaizhu Huang, Rong Jin, Zenglin Xu, and Cheng-Lin Liu. Robust metric learning by smooth optimization. In *UAI*, pages 244–251, 2010.
- Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *ICML*, pages 457–464, 2009.
- Thorsten Joachims. Making large-scale support vector machine learning practical. pages 169–184, 1999.
- Thorsten Joachims. Training linear svms in linear time. In *KDD*, pages 217–226, 2006.
- Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.
- Guanghui Lan. An optimal method for stochastic composite optimization. *Math. Program.*, 2010.
- Guanghui Lan, Zhaosong Lu, and Renato D. C. Monteiro. Primal-dual first-order methods with $1/\epsilon$ iteration-complexity for cone programming. *Math. Program.*, 126:1–29, January 2011.
- Qihang Lin. A smoothing stochastic gradient method for composite optimization. *ArXiv e-prints*, 2010.
- Jun Liu, Shuiwang Ji, and Jieping Ye. Multi-task feature learning via efficient l_2 , l_1 -norm minimization. In *UAI*, pages 339–348, 2009.

- Sofia Mosci, Silvia Villa, Alessandro Verri, and Lorenzo Rosasco. A primal-dual algorithm for group sparse regularization with overlapping groups. In *NIPS*, pages 2604–2612, 2010.
- Arkadi Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM J. on Optimization*, 15:229–251, 2005.
- Yu Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103:127–152, 2005a.
- Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM J. on Optimization*, pages 235–249, 2005b.
- Yu. Nesterov. Gradient methods for minimizing composite objective function. Core discussion papers, 2007.
- Feiping Nie, Heng Huang, Xiao Cai, and Chris Ding. Efficient and robust feature selection via joint 2,1-Norms minimization. In *Advances in Neural Information Processing Systems 23*, pages 1813–1821. 2010.
- John C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208, Cambridge, MA, 1998.
- Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell. An efficient projection for l1, infinity regularization. In *ICML*, pages 857–864, 2009.
- Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed Minimum-Rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.*, 52:471–501, August 2010.
- Jasson D. M. Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on Machine learning*, pages 713–719, 2005.
- Lorenzo Rosasco, Ernesto De Vito, Andrea Caponnetto, Michele Piana, and Alessandro Verri. Are loss functions all the same? *Neural Comput.*, 16:1063–1076, 2004.
- Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, pages 807–814, 2007.
- Steve Smale and Ding-Xuan Zhou. Estimating the approximation error in learning theory. *Anal. Appl. (Singap.)*, 1(1):17–41, 2003.
- Alex J. Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14:199–222, 2004.
- N. Srebro, J. D. M. Rennie, and T. S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2005.

- Liang Sun, Jun Liu, Jianhui Chen, and Jieping Ye. Efficient recovery of jointly sparse vectors. In *Advances in Neural Information Processing Systems 22*, pages 1812–1820, 2009.
- J. F. Traub, G. W. Wasilkowski, and H. Woźniakowski. *Information-based complexity*. Academic Press Professional, Inc., 1988.
- Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, 2008.
- V.N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.
- Qiang Wu and Ding-Xuan Zhou. Svm soft margin classifiers: Linear programming versus quadratic programming. *Neural Comput.*, 17:1160–1187, 2005.
- Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *NIPS*, pages 2116–2124. 2009.
- Haiqin Yang, Zenglin Xu, Irwin King, and Michael R. Lyu. Online learning for group lasso. In *ICML*, pages 1191–1198, 2010.
- Gene Yeo and Christopher B. Burge. Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. In *RECOMB*, pages 322–331, 2003.
- Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *JRSS*, 68:49–67, 2006.
- Tianyi Zhou, Dacheng Tao, and Xindong Wu. Nesvm: a fast gradient method for support vector machines. *CoRR*, abs/1008.4000, 2010a.
- Yang Zhou, Rong Jin, and Steven C.H. Hoi. Exclusive lasso for multi-task feature selection. In *AISTAT*, pages 988–995, 2010b.
- Ji Zhu, Saharon Rosset, Trevor Hastie, and Rob Tibshirani. 1-norm support vector machines. In *NIPS*, 2003.
- M. Zhu and T. Chan. An efficient primal-dual hybrid gradient algorithm for total variation image restoration. *UCLA CAM Report*, pages 08–34, 2008.

Appendix A. Derivation of constant c for (generalized) hinge loss

As mentioned before, it is easy to derive the constant c in equations (6) and (7) for the non-smooth loss functions listed before under the assumption that $\|\mathbf{x}\|_2 \leq 1$. As an example, here we derive the constant for hinge loss and generalized hinge loss. For other non-smooth loss functions, we can derive the value of c in a similar way. For hinge loss, $L(\mathbf{w}, \boldsymbol{\alpha})$ in (5) is given by

$$L(\mathbf{w}, \boldsymbol{\alpha}; \mathbf{X}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \alpha_i (1 - y_i \mathbf{w}^\top \mathbf{x}_i),$$

and its partial gradients are

$$\begin{aligned} G_{\alpha}(\mathbf{w}, \alpha) &= \frac{1}{n} \mathbf{1} - \frac{1}{n} (\mathbf{x}_1 y_1, \dots, \mathbf{x}_n y_n)^\top \mathbf{w}, \\ G_{\mathbf{w}}(\mathbf{w}, \alpha) &= -\frac{1}{n} \mathbf{X}(\alpha \circ \mathbf{y}), \end{aligned}$$

where $\mathbf{1}$ denotes a vector of all ones, and \circ denotes the element-wise product. Then,

$$\begin{aligned} \|G_{\alpha}(\mathbf{w}_1, \alpha_1) - G_{\alpha}(\mathbf{w}_2, \alpha_2)\|_2^2 &= \frac{1}{n^2} \sum_{i=1}^n (\mathbf{w}_1^\top \mathbf{x}_i y_i - \mathbf{w}_2^\top \mathbf{x}_i y_i)^2 \leq \frac{1}{n} \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \\ \|G_{\mathbf{w}}(\mathbf{w}_1, \alpha_1) - G_{\mathbf{w}}(\mathbf{w}_2, \alpha_2)\|_2^2 &= \frac{1}{n^2} \left\| \sum_{i=1}^n (\alpha_i^1 - \alpha_i^2) y_i \mathbf{x}_i \right\|_2^2 \leq \frac{1}{n} \sum_{i=1}^n (\alpha_i^1 - \alpha_i^2)^2 = \frac{1}{n} \|\alpha_1 - \alpha_2\|_2^2, \end{aligned}$$

which implies $c = 1/n$. For the example of generalized hinge loss, $L(\mathbf{w}, \alpha)$ in (5) is

$$L(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^n \alpha_i^1 (1 - a y_i \mathbf{w}^\top \mathbf{x}_i) + \alpha_i^2 (1 - y_i \mathbf{w}^\top \mathbf{x}_i),$$

where $\alpha = [\alpha^1, \alpha^2] \in \mathcal{Q}_\alpha = \{\alpha : \alpha \in [0, 1]^{n \times 2}, \alpha^1 + \alpha^2 \leq \mathbf{1}\}$, and its partial gradients are

$$\begin{aligned} G_{\alpha}(\mathbf{w}, \alpha) &= \frac{1}{n} [\mathbf{1}, \mathbf{1}] - \frac{1}{n} [a(\mathbf{x}_1 y_1, \dots, \mathbf{x}_n y_n)^\top \mathbf{w}, (\mathbf{x}_1 y_1, \dots, \mathbf{x}_n y_n)^\top \mathbf{w}], \\ G_{\mathbf{w}}(\mathbf{w}, \alpha) &= -\frac{1}{n} \mathbf{X}(a(\alpha^1 \circ \mathbf{y}) + \alpha^2 \circ \mathbf{y}), \end{aligned}$$

where $\mathbf{1}$ denotes a vector of all ones, and \circ denotes the element-wise product. Then for any $\mathbf{w}_1, \mathbf{w}_2$ and $\alpha_1 = (\alpha^{1,1}, \alpha^{2,1}), \alpha_2 = (\alpha^{1,2}, \alpha^{2,2}) \in \mathcal{Q}_\alpha$, given $\|\mathbf{x}\|_2 \leq 1$, we have

$$\begin{aligned} \|G_{\alpha}(\mathbf{w}_1, \alpha_1) - G_{\alpha}(\mathbf{w}_2, \alpha_2)\|_F^2 &= \frac{a^2 + 1}{n^2} \sum_{i=1}^n (\mathbf{w}_1^\top \mathbf{x}_i y_i - \mathbf{w}_2^\top \mathbf{x}_i y_i)^2 \leq \frac{a^2 + 1}{n} \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \\ \|G_{\mathbf{w}}(\mathbf{w}_1, \alpha_1) - G_{\mathbf{w}}(\mathbf{w}_2, \alpha_2)\|_2^2 &= \frac{1}{n^2} \left\| \sum_{i=1}^n a(\alpha_i^{1,1} - \alpha_i^{1,2}) y_i \mathbf{x}_i + \sum_{i=1}^n (\alpha_i^{2,1} - \alpha_i^{2,2}) y_i \mathbf{x}_i \right\|_2^2 \\ &\leq \frac{2a^2}{n} \sum_{i=1}^n (\alpha_i^{1,1} - \alpha_i^{1,2})^2 + \frac{2}{n} \sum_{i=1}^n (\alpha_i^{2,1} - \alpha_i^{2,2})^2 \\ &\leq \frac{2a^2}{n} \|\alpha_1 - \alpha_2\|_F^2, \end{aligned}$$

which implies $c = (a^2 + 1)/n$ in equation (6) and $c = (2a^2)/n$ in equation (7). We can derive the value of c in (6) and (7) similarly for other non-smooth loss functions.

Appendix B. Proof of Lemma 3

Since

$$\begin{aligned} G_{\alpha}(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y}) &= \mathbf{a}(\mathbf{X}, \mathbf{y}) + H(\mathbf{X}, \mathbf{y})^\top \mathbf{w}, \\ G_{\mathbf{w}}f(\mathbf{w}, \alpha; \mathbf{X}, \mathbf{y}) &= \mathbf{b}(\mathbf{X}, \mathbf{y}) + H(\mathbf{X}, \mathbf{y}) \alpha. \end{aligned}$$

Then

$$\begin{aligned}\|G_{\mathbf{w}}(\mathbf{w}_1, \boldsymbol{\alpha}_1; \mathbf{X}, \mathbf{y}) - G_{\mathbf{w}}(\mathbf{w}_2, \boldsymbol{\alpha}_2; \mathbf{X}, \mathbf{y})\|_2^2 &\leq \|H(\mathbf{X}, \mathbf{y})^\top (\mathbf{w}_1 - \mathbf{w}_2)\|_2^2 \leq c \|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \\ \|G_{\mathbf{w}}(\mathbf{w}_1, \boldsymbol{\alpha}_1; \mathbf{X}, \mathbf{y}) - G_{\mathbf{w}}(\mathbf{w}_2, \boldsymbol{\alpha}_2; \mathbf{X}, \mathbf{y})\|_2^2 &\leq \|H(\mathbf{X}, \mathbf{y})(\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2)\|_2^2 \leq c \|\boldsymbol{\alpha}_1 - \boldsymbol{\alpha}_2\|_2^2,\end{aligned}$$

where we use the assumption $\|H(\mathbf{X}, \mathbf{y})\|_2^2 = \|H(\mathbf{X}, \mathbf{y})^\top\|_2^2 \leq c$.

Appendix C. Proof of Lemma 6

In order to proof Lemma 6, we first present the following lemma with its proof.

Lemma 12 *Let Z be a convex compact set, and $U \subseteq Z$ be convex and closed, $\mathbf{z}_0 \in Z$, and $\gamma > 0$. Considering the points,*

$$\begin{aligned}\mathbf{z}_h &= \arg \min_{\mathbf{z} \in U} \frac{1}{2} \|\mathbf{z} - (\mathbf{z}_0 - \gamma \xi)\|_2^2, \\ \mathbf{z}_1 &= \arg \min_{\mathbf{z} \in U} \frac{1}{2} \|\mathbf{z} - (\mathbf{z}_0 - \gamma \eta)\|_2^2,\end{aligned}$$

then for any $\mathbf{z} \in U$, we have

$$\gamma \eta^\top (\mathbf{z}_h - \mathbf{z}) \leq \frac{1}{2} \|\mathbf{z} - \mathbf{z}_0\|_2^2 - \frac{1}{2} \|\mathbf{z} - \mathbf{z}_1\|_2^2 + \gamma^2 \|\xi - \eta\|_2^2 - \frac{1}{2} [\|\mathbf{z}_h - \mathbf{z}_0\|_2^2 + \|\mathbf{z}_h - \mathbf{z}_1\|_2^2].$$

Equipped with above lemma, it is straightforward to prove Lemma 6. We note that the two updates in Lemma 5 are the same as the two updates in Lemma 12 if we make the following correspondences:

$$\begin{aligned}U &= Z = \mathbb{R}^d \times \mathcal{Q}_\alpha, \quad \mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \alpha \end{pmatrix} \in U, \\ \mathbf{z}_0 &= \begin{pmatrix} \mathbf{u}_{t-1} \\ \beta_{t-1} \end{pmatrix}, \quad \mathbf{z}_h = \begin{pmatrix} \mathbf{w}_t \\ \alpha_t \end{pmatrix}, \quad \mathbf{z}_1 = \begin{pmatrix} \mathbf{u}_t \\ \beta_t \end{pmatrix}, \\ \xi &= \begin{pmatrix} G_{\mathbf{w}}(\mathbf{u}_{t-1}, \alpha_t) + \lambda \mathbf{v}_t \\ -G_\alpha(\mathbf{u}_{t-1}, \beta_{t-1}) \end{pmatrix}, \quad \eta = \begin{pmatrix} G_{\mathbf{w}}(\mathbf{w}_t, \alpha_t) + \lambda \mathbf{v}_t \\ -G_\alpha(\mathbf{w}_t, \alpha_t) \end{pmatrix}.\end{aligned}$$

Then the inequality in Lemma 6 follows immediately the inequality in Lemma 12.

Lemma 12 is a special case of Lemma 3.1 (Nemirovski, 2005) for Euclidean norm. A proof is provided here for completeness.

Proof [of Lemma 12] Since

$$\begin{aligned}\mathbf{z}_h &= \arg \min_{\mathbf{z} \in U} \frac{1}{2} \|\mathbf{z} - (\mathbf{z}_0 - \gamma \xi)\|_2^2, \\ \mathbf{z}_1 &= \arg \min_{\mathbf{z} \in U} \frac{1}{2} \|\mathbf{z} - (\mathbf{z}_0 - \gamma \eta)\|_2^2,\end{aligned}$$

by the first order optimality condition, we have

$$(\mathbf{z} - \mathbf{z}_h)^\top (\gamma \xi - \mathbf{z}_0 + \mathbf{z}_h) \geq 0, \forall \mathbf{z} \in U, \quad (19)$$

$$(\mathbf{z} - \mathbf{z}_1)^\top (\gamma \eta - \mathbf{z}_0 + \mathbf{z}_1) \geq 0, \forall \mathbf{z} \in U. \quad (20)$$

Applying (19) with $\mathbf{z} = \mathbf{z}_1$ and (20) with $\mathbf{z} = \mathbf{z}_h$, we get

$$\begin{aligned}\gamma(\mathbf{z}_h - \mathbf{z}_1)^\top \xi &\leq (\mathbf{z}_0 - \mathbf{z}_h)^\top (\mathbf{z}_h - \mathbf{z}_1), \\ \gamma(\mathbf{z}_1 - \mathbf{z}_h)^\top \eta &\leq (\mathbf{z}_0 - \mathbf{z}_1)^\top (\mathbf{z}_1 - \mathbf{z}_h).\end{aligned}$$

Summing up the two inequalities, we have

$$\gamma(\mathbf{z}_h - \mathbf{z}_1)^\top (\xi - \eta) \leq (\mathbf{z}_1 - \mathbf{z}_h)^\top (\mathbf{z}_h - \mathbf{z}_1) = -\|\mathbf{z}_1 - \mathbf{z}_h\|_2^2.$$

Then

$$\gamma\|\xi - \eta\|_2\|\mathbf{z}_h - \mathbf{z}_1\|_2 \geq -\gamma(\mathbf{z}_h - \mathbf{z}_1)^\top (\xi - \eta) \geq \|\mathbf{z}_1 - \mathbf{z}_h\|_2^2. \quad (21)$$

We continue the proof as follows:

$$\begin{aligned}\frac{1}{2}\|\mathbf{z} - \mathbf{z}_0\|_2^2 - \frac{1}{2}\|\mathbf{z} - \mathbf{z}_1\|_2^2 &= \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 + (\mathbf{z} - \mathbf{z}_1)^\top (\mathbf{z}_1 - \mathbf{z}_0) \\ &= \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 + (\mathbf{z} - \mathbf{z}_1)^\top (\gamma\eta + \mathbf{z}_1 - \mathbf{z}_0) - (\mathbf{z} - \mathbf{z}_1)^\top \gamma\eta \\ &\geq \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z} - \mathbf{z}_1)^\top \gamma\eta \\ &= \underbrace{\frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma\eta + (\mathbf{z}_h - \mathbf{z})^\top \gamma\eta}_{\epsilon},\end{aligned}$$

where the inequality follows from (20).

$$\begin{aligned}\epsilon &= \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma\eta \\ &= \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma(\eta - \xi) - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma\xi \\ &= \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma(\eta - \xi) \\ &\quad + (\mathbf{z}_1 - \mathbf{z}_h)^\top (\gamma\xi - \mathbf{z}_0 + \mathbf{z}_h) - (\mathbf{z}_1 - \mathbf{z}_h)^\top (\mathbf{z}_h - \mathbf{z}_0) \\ &\geq \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma(\eta - \xi) - (\mathbf{z}_1 - \mathbf{z}_h)^\top (\mathbf{z}_h - \mathbf{z}_0) \\ &= \frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_h - \mathbf{z}_0)^\top \mathbf{z}_0 - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma(\eta - \xi) - (\mathbf{z}_1 - \mathbf{z}_h)^\top \mathbf{z}_h \\ &= \left[\frac{1}{2}\|\mathbf{z}_1\|_2^2 - \frac{1}{2}\|\mathbf{z}_h\|_2^2 - (\mathbf{z}_1 - \mathbf{z}_h)^\top \mathbf{z}_h \right] + \left[\frac{1}{2}\|\mathbf{z}_h\|_2^2 - \frac{1}{2}\|\mathbf{z}_0\|_2^2 - (\mathbf{z}_h - \mathbf{z}_0)^\top \mathbf{z}_0 \right] - (\mathbf{z}_h - \mathbf{z}_1)^\top \gamma(\eta - \xi) \\ &= \frac{1}{2}\|\mathbf{z}_h - \mathbf{z}_1\|_2^2 + \frac{1}{2}\|\mathbf{z}_h - \mathbf{z}_0\|_2^2 - \gamma\|\mathbf{z}_h - \mathbf{z}_1\|_2\|\eta - \xi\|_2 \\ &\geq \frac{1}{2}\{\|\mathbf{z}_h - \mathbf{z}_1\|^2 + \|\mathbf{z}_h - \mathbf{z}_0\|^2\} - \gamma^2\|\eta - \xi\|_2^2,\end{aligned}$$

where the first inequality follows from (19), and the last inequality follows from (21). Combining the above results, we have

$$\gamma(\mathbf{z}_h - \mathbf{z})^\top \eta \leq \frac{1}{2}\|\mathbf{z} - \mathbf{z}_0\|_2^2 - \frac{1}{2}\|\mathbf{z} - \mathbf{z}_1\|_2^2 + \gamma^2\|\eta - \xi\|_2^2 - \frac{1}{2}\{\|\mathbf{z}_h - \mathbf{z}_1\|_2^2 + \|\mathbf{z}_h - \mathbf{z}_0\|_2^2\}.$$

■

Appendix D. Proof of Lemma 10

By introducing Lagrangian multiplier for constraint $\sum_i \alpha_i v_i \leq \rho$, we have the following min-max problem

$$\max_{\eta} \min_{\alpha \in [0,1]^n} \frac{1}{2} \|\alpha - \hat{\alpha}\|^2 + \eta \left(\sum_i \alpha_i v_i - \rho \right).$$

The solution to α is $\alpha_i = [\hat{\alpha}_i - \eta^* v_i]_{[0,1]}$. By KKT condition, the optimal η^* is equal to 0 if $\sum_i [\hat{\alpha}_i]_{[0,1]} v_i < \rho$, otherwise we have

$$\sum_i [\hat{\alpha}_i - \eta^* v_i]_{[0,1]} v_i - \rho = 0.$$

Since the left side of above equation is a monotonically decreasing function in η^* , we can compute η^* by efficient bi-section search.

Appendix E. Proof of Lemma 11

Using the convex conjugate $V_*(\eta)$ of $V(z)$, the composite mapping can be written as

$$\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \hat{\mathbf{w}}\|_2^2 + \lambda \max_{\eta} (\eta \|\mathbf{w}\| - V_*(\eta)).$$

The problem is equivalent to maximize the following function on η ,

$$\left(\min_{\mathbf{w}} \frac{1}{2} \|\mathbf{w} - \hat{\mathbf{w}}\|_2^2 + \lambda \eta \|\mathbf{w}\| \right) - \lambda V_*(\eta).$$

Let $\mathbf{w}(\eta)$ denote the solution to the minimization problem. Then the optimal solution of η satisfies

$$\lambda \|\mathbf{w}(\eta)\| - \lambda V'_*(\eta) = 0,$$

i.e.

$$\|\mathbf{w}(\eta)\| - V'_*(\eta) = 0.$$

It is easy to show that $\|\mathbf{w}(\eta)\|$ is a non-increasing function in η . Similarly, since $V_*(\eta)$ is a convex function, its negative gradient $-V'_*(\eta)$ is a non-increasing function. Therefore, we can compute the optimal η by bi-section search.